

From risk to alerts: Statistical methods across changing questions and data structures



Doctoral thesis
for the award of the doctoral degree
of the Faculty of Mathematics and Natural Sciences
of the University of Cologne

submitted by

Lucia Ameis

accepted in the year 2026

Acknowledgments

Foremost, I would like to thank my colleagues who have accompanied me on this journey. Special thanks go to my supervisor, Kathrin Möllenhoff, who has supported me and consistently given me the opportunity to grow. I would like to thank Holger Schwender, who served as my second reviewer. I would also like to thank Niklas and the other members of GRK 2624 for their support.

Most of all, I'd like to thank my mother Ulrike, my husband Paul, my aunt Annette, and the rest of my family and friends, who have always been there for me. Also, thanks to my cats for being cute and cuddly!

Danke!

Abstract

As biostatistical research becomes more specialized and diverse, a careful alignment between research question, data structure, and methodology is required for valid statistical conclusions. This necessitates the development of flexible approaches that are simultaneously tailored to the research question, while accounting for the underlying data structure. This thesis comprises three contributions that discuss the interplay between data, research questions, and methodology for two major thematic areas: risk estimation for time-to-event data, and alert identification for continuous covariates.

The first contribution introduces a nonparametric estimator for the relative risk of two groups to experience the event. By directly estimating the relative risk, the proposed method avoids problems that arise from the misinterpretation of the odds ratio or the hazard ratio as the relative risk, which is often found in practice. Therefore, it is better aligned with the research question of primary interest. The procedure utilizes the Kaplan-Meier estimator. In contrast to an existing approach, it thus avoids parametric assumptions and merely assumes proportional risks, that is, a constant risk ratio over time. As a result, it remains mostly interpretable even when this assumption is violated and is applicable to a broader spectrum of data.

The second and third contributions introduce model-based approaches for the identification of alerts, at which a pre-specified threshold of the response variable is exceeded. Both methods use hypothesis tests, enabling inference based on statistical significance. The second contribution extends existing methods by shifting from the identification of the onset of an effect to the identification of entire covariate regions, such as time intervals, over which the response changes. This approach retains the advantages of parametric model-based methods, including interpolation between observed covariate values and applicability to a broad range of data structures. This is achieved by basing the analysis on the first derivative of the parametric model. The third contribution extends model-based alert identification to two continuous covariates and generalizes naturally to higher-dimensional settings. The proposed approach enables the identification of an alert for one covariate while the other is fixed, or alternatively, an alert relationship between the covariates, depending on the research question. To accommodate the increased complexity of multidimensional data, the approach incorporates the flexible framework of generalized additive models for location, scale, and shape.

Each method is evaluated by means of a simulation study, and its practical relevance is investigated by case studies from diabetes research or toxicology. Collectively, these contributions provide methodological advances that improve the alignment between research questions and statistical analysis, while accounting for the underlying data structure.

Contents

Acknowledgments	II
Abstract	III
Notation	V
Abbreviations	IX
1 Introduction	1
2 Preliminaries	5
2.1 Time-to-event analysis	5
2.2 Model-based identification of alerts	10
2.2.1 Observation-based alerts	10
2.2.2 Dose-response modeling	11
2.2.3 Model-based alerts	13
2.2.4 The bootstrap	18
2.2.5 Estimation of the lower simultaneous confidence band	20
2.2.6 Model-based alerts in a multidimensional setting	24
2.3 Regression modeling	27
2.3.1 The generalized linear model	28
2.3.2 The generalized additive model	29
2.3.3 Generalized additive models for location, scale, and shape	31
3 Articles	36
3.1 A nonparametric proportional risk model to assess a treatment effect in time-to-event data	37
3.2 Identification of changes in gene expression	88
3.3 Estimating effect thresholds and beyond: A flexible framework for multi- variate alert detection	135
4 Discussion	177
References	180
Further publications	188

Notation

In general	
$\mathbb{R}$	Real numbers
$\mathbb{P}(X \leq a)$	Probability of random variable X being less than or equal to a
$\mathbb{E}$	Expected value
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
In the context of time-to-event analysis	
X	Random variable describing the time until an event of interest
f	Probability density function
F	Cumulative distribution function
h	Hazard rate
h_{base}	Baseline hazard rate of Cox's proportional hazards model (see (1))
S	Survival function
S_{base}	Baseline survival function of Cox's proportional hazards model
L	Likelihood
D	Number of distinct event time points
$t_i, i = 1, \dots, D$	Event time points
$Y_i, i = 1, \dots, D$	Number of individuals at risk at t_i
$d_i, i = 1, \dots, D$	Number of events at t_i
$c_{[i-1, i]}, i = 2, \dots, D$	Number of censored observations between t_{i-1} and t_i
$HR(t)$	Hazard ratio depending on time
$OR(t)$	(Failure) Odds ratio depending on time
$RR(t)$	Relative risk depending on time
$\beta \mathbf{X}$	Linear effect of covariates
$\mathbf{x}, \tilde{\mathbf{x}}$	Realization of expression vectors
In the context of model-based alert identification	
Note that the running index of the dose levels in dose-response models starts with 0 to reflect the placebo group. This is different to the running index in the preliminary section on regression modeling and the article presented in Section 3.2.	
$y_{i,j}$	j th observation in dose group i (see (2))
D	Dose region (example covariate)
m	Number of dose levels

n_i	Number of observations at the i th dose level
n	Number of total observations
d_i	Measured dose groups, d_0 indicates the placebo (see (2))
f	Parametric dose-response regression model or B-splines (see (2))
θ	Parameter vector of f (see (2))
$\epsilon_{i,j}$	Random error corresponding to the j th observation in dose group i ; often assumed independent identically normally distributed with mean 0 and standard deviation σ (see (2))
κ	Knots that partition the covariate region into intervals; used for fitting splines
λ	Pre-specified threshold
$L^\alpha(d, \hat{\theta})$	Lower simultaneous confidence band of $ f(d, \theta) - f(d_0, \theta) $ (see (7))
$\hat{\sigma}_{ f(d, \theta) - f(d_0, \theta) }$	Estimated standard error of $ f(d, \theta) - f(d_0, \theta) $ (see (7))
$Z(d, \hat{\theta}, \theta)$	Standardized value of $ f(d, \theta) - f(d_0, \theta) $ (see (9))
$y_{i,j}^*, \epsilon_{i,j}^*, \hat{\theta}^*$	Bootstrapped observation, error and estimated parameter vector of the first-level bootstrap
$y_{i,j}^{**}, \epsilon_{i,j}^{**}, \hat{\theta}^{**}$	Bootstrapped observation, error and estimated parameter vector of the second-level bootstrap
B_1	Number of first-level bootstrap samples
B_2	Number of second-level bootstrap samples
D^*	$\max_{d \in D} Z(d, \hat{\theta}^*, \hat{\theta})$
c	$(1 - \alpha)$ -quantile value of $\max_{d \in D} Z(d, \hat{\theta}, \theta)$ (see (7))

	In the context of general bootstraps
F	Cumulative distribution function
$\mathbf{x}$	Observed sample
$\mathbf{x}^*$	Generated bootstrap sample
$\hat{\phi}$	Estimated parameter of interest
B	Number of bootstrap samples
$\hat{\phi}^*$	Estimated parameter from a bootstrap sample
$\sigma_{\hat{\phi}}$	Standard error of $\hat{\phi}$
z_α	α -quantile, bootstrapped or from the normal distribution or the Student t -distribution (see (5))
$Z(\hat{\phi}^*)$	Standardized value of $\hat{\phi}^*$ (see (6))

$\hat{\phi}_\alpha^*$ Empirical α -quantile from the bootstrapped set $\hat{\phi}_l^*$, $l = 1, \dots, B$

In the context of regression modeling	
Y	Random variable describing the response
$\mathbf{Y} = (Y_1, \dots, Y_m)$	Vector of random variables describing the response; Y_i , $i = 1, \dots, m$, are usually assumed to be independent
y	Realization of Y
$\mathbf{y} = (y_1, \dots, y_m)$	Realization of $\mathbf{Y}$
X	Random variable describing the covariates
$\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$	Realization of the covariate vector corresponding to the response y_i
f	General regression model
ϵ	Error vector; often assumed (multivariate) normally distributed with mean $\mathbf{0}$ and covariance matrix Σ
g, h	Link function and response function
$\eta(\mathbf{x})$	Predictor term depending on covariate vector $\mathbf{x}$ (see (10))
f_k	Additive effect in the predictor term of GAMs/GAMLSS
$\boldsymbol{\eta}$	Predictor term in matrix notation (see (11))
$\beta\mathbf{X}$	Covariate matrix and regression coefficients, also linear additive effect
B_r	Spline basis function
$\mathbf{V}_k\boldsymbol{\gamma}_k$	Additive effect in matrix notation with penalization
λ_j	Smoothing parameter corresponding to j th additive effect; not to be confused with pre-specified threshold λ
$\mathbf{K}_j$	Penalization matrix corresponding to j th additive effect
l_{pen}	Penalized log-likelihood (see (12) for GAM and (15))
LS_{pen}	Penalized least squares (see (13))
$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\kappa})$	Distribution with four distribution parameters
Specific to Section 3.1	
The notation is mainly set in the context of time-to-event analysis. In addition:	
β	Logarithmic relative risk
$T, t_{\min}, t_{\max}$	Theoretical time frame the analysis is restricted to and its minimum and maximum
$\tilde{T}_i, i \in \{0, 1\}$	Ordered sets of observed event time points per group
$\tilde{T}, \tilde{t}_{\min}, \tilde{t}_{\max}$	Observed time frame during which events occur in both groups and its minimum and maximum

$\widehat{\text{var}}(\hat{S}(t))$	Greenwood's formula for the variance of the Kaplan-Meier estimator
$\hat{\beta}_t$	Estimated logarithmic relative risk at t
$\omega(t)$	Estimated variance of $\hat{\beta}_t$
$RD(t)$	Risk difference at t
$NNT(t)$	Number needed to treat at t

Specific to Section 3.2

The notation is mainly set in the context of model-based alert identification. Note that the indices differ. In addition:

T	Study period
$\mathcal{L}_{\mathbf{y}, \mathbf{t}}(\boldsymbol{\theta})$	Likelihood
f'	First derivative of parametric model or spline model
$L^\alpha(t, \hat{\boldsymbol{\theta}})$	Lower simultaneous confidence band of f'
$\tilde{S}$	Time period of significant change in response
$(S_1, \dots, S_s)$	Partition of $\tilde{S}$ into coherent subsets

Specific to Section 3.3

The notation combines the context of model-based alert identification and regression modeling. Note that the indices differ. Note that the running indices of the covariates start at 0 to reflect reference levels. In addition:

T, D	Study period and dose range as representation for two continuous covariates
$\boldsymbol{\theta}$	Parameter vector of the regression model of the mean
$\boldsymbol{\vartheta}$	Parameter vector of the regression model of the standard deviation
$LB^\alpha(\tilde{t}, d, \boldsymbol{\theta}), UB^\alpha(\tilde{t}, d, \boldsymbol{\theta})$	Lower/upper simultaneous confidence band for a fixed time point $\tilde{t}$
$LP^\alpha(t, d, \boldsymbol{\theta}), UP^\alpha(t, d, \boldsymbol{\theta})$	Lower/upper simultaneous confidence plane
$\Delta(t, d, \hat{\boldsymbol{\theta}})$	Representative of the curve or plane the confidence region is constructed for; defined for notational simplicity.

Abbreviations

AL(O)EC	Absolute lowest (observed) effective concentration
CB/CI	Simultaneous confidence band/Confidence interval
CDF	Cumulative distribution function
EU	Exponentiated-uniform
GAM	Generalized additive model
GAMLSS	Generalized additive models for location, scale, and shape
GLM	Generalized linear model
GO	Gene ontology
HR	Hazard ratio
IQR	Interquartile range
IWLS	Iterative weighted least squares
L(O)EC	Lowest (observed) effective concentration
LL	Log-logistic
MEC	Multivariate effective concentration (also MED)
MED	Minimum effective dose (Preliminaries) or multivariate effective dose (Section 3.3)
MSE	Mean squared error
NAFLD	Non-alcoholic fatty liver disease
NNT	Number needed to treat
NOEC	No observed effect concentration
NPPR	Nonparametric proportional risks
OLS	Ordinary least squares
OR	Odds ratio
PH	Proportional hazards
PO	Proportional odds
PPR	Parametric porportional risks
PR	Proportional risks
PWLS	Penalized weighted least squares
RCT	Randomized controlled trial
RD	Risk difference
RJG	Rest-and-jump gene
RMSE	Root mean squared error
RR	Relative risk
SD	Standard diet
WD	Western diet
WLS	Weighted least squares

1 Introduction

Modern biostatistical research is characterized by growing data complexity (Loh 2021; Lee et al. 2019) and an increasing diversity of biomedical research questions, as reflected by the expanding landscape of scientific publications (González-Márquez et al. 2024). As a result, valid statistical conclusions require methods that account for both the data structure and the research question of interest (Mishra et al. 2019; Sauerbrei et al. 2026). In this dissertation, the interplay between data structure, research question, and methodology is addressed in two thematic areas: the estimation of *risks in time-to-event data* and the *model-based identification of alerts* for continuous covariates.

This interplay can take several directions, of which two are discussed in this thesis. The first direction concerns aligning the research question with the method. The second direction regards the assumptions underlying statistical methods and their implications for applicability. On the most basic level, these are the assumptions that are essential to the method. Relaxing such assumptions can substantially broaden the range of data structures to which a method can be applied. With regard to regression-based approaches, flexible procedures that do not impose restrictive assumptions on the candidate models can be applied to a wider range of data structures than more limiting methods. More general modeling frameworks, such as *generalized additive models for location, scale, and shape* (GAMLSS; Rigby et al. 2005), further extend this flexibility and enable the analysis of complex data, including in multidimensional settings.

These considerations motivate the discussion of statistical methodology and the development of adaptations or new approaches in this thesis. The role of the research question is examined in both thematic areas. In the context of risk estimation for time-to-event data, the focus additionally lies on avoiding restrictive parametric assumptions. For model-based alert identification, the integration of more complex types of regression frameworks is discussed.

The first thematic area discussed in this thesis concerns the analysis of time-to-event data, in which the time until a specific event occurs is measured, for example, the time until death or worsening of an illness (Kleinbaum et al. 2012). An essential task of this analysis involves a comparison of the treatment groups in terms of risks. Widespread models include *Cox's proportional hazards* model (Cox 1972) and the *log-logistic proportional odds* model (Kleinbaum et al. 2012) that quantify group differences via the *hazard ratio* (HR) or the *odds ratio* (OR), respectively. Both measures are often misinterpreted as *relative risk* (RR), which is the ratio of the *cumulative distribution functions* (CDF). However, although HR, OR, and RR typically indicate the same direction of an effect, they are not identical (Sutradhar et al. 2018; Knol et al. 2011). As a result, commonly used models not fully align with the research question of by what factor the risk for one group is higher than for the other. Additionally, especially the assumption of *proportional hazards* (PH) is rarely assessed in practice (Jachno et al. 2019), putting the validity of the results into

question. Similarly, the assumption of *proportional odds* (PO) is not always suitable. To allow for the direct estimation of the RR, thus aligning research question and methodology, Kuss et al. (2021) introduced a parametric approach for a two-group situation based on the alternative assumption of *proportional risks* (PR). However, the approach additionally assumes that the data follow the *exponentiated-uniform* distribution.

Building on the ideas of PR and the direct estimation of the RR, this dissertation aims to propose a nonparametric procedure that is applicable to a broader spectrum of data, since it does not rely on a restrictive parametric assumption. This is achieved by the introduction of an estimator based on a weighted mean of the ratio of two *Kaplan-Meier estimators* (Kaplan et al. 1958) evaluated at different time points. Even if the assumption of PR is violated, the method allows for a meaningful interpretation of the RR in most cases, as it estimates the mean RR over time. Utilizing a *percentile bootstrap* (Efron 1981), a confidence interval for the RR can be derived. Due to the duality between hypothesis test and confidence intervals, this provides a test for group differences. In addition to a relative measure such as the RR, an absolute measure should always be reported (The Academy of Medical Sciences 2017). Therefore, an estimator is derived for the number needed to treat, which allows for a more holistic interpretation.

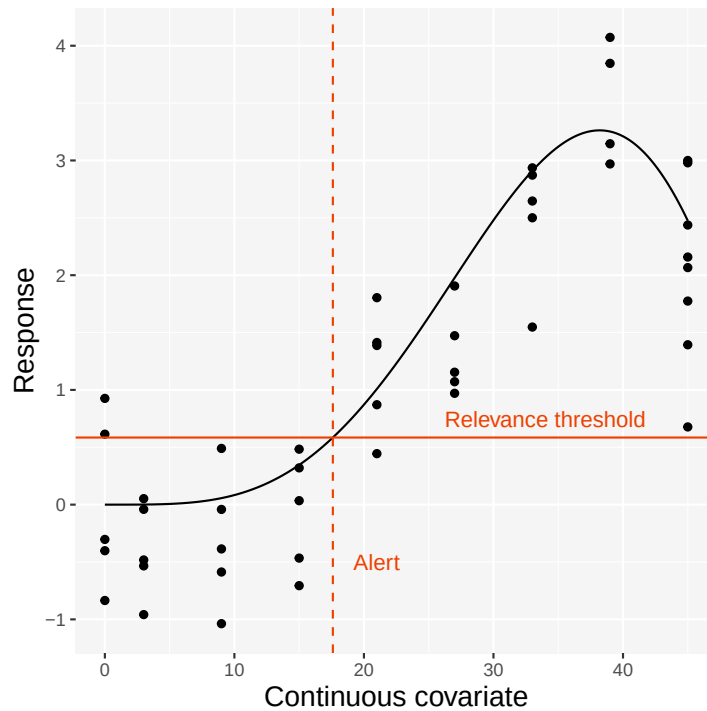


Figure 1: Visualization of the basic idea model-based of alert identification. The black dots indicate the individual measurements. The black line indicates the estimated response. An alert is identified when the estimated response exceeds a pre-specified relevance threshold.

The second thematic area of this dissertation discusses the identification of alerts for continuous covariates. Alert concentrations such as the lowest concentration with a noticeable effect (Jensen et al. 2019) or the EC_{50} value, where half of the maximum predicted effect is observed (Bretz et al. 2010), are an important tool for the analysis of dose-response or

time-response data, especially in toxicological research (Kappenberg et al. 2023). Figure 1 provides a basic visualization of the concept of alert identification. Traditional approaches, including multiple t -tests or the recommended Dunnett test (Hothorn 2014), are limited to comparisons at observed dose levels and do not allow for interpolation between those. In response, Kappenberg et al. (2021) introduced a parametric approach that assumes a monotonic response and is constructed around the $4pLL$ model, otherwise known as the *sigmoid emax* model (Bornkamp et al. 2009). The procedure is based on a hypothesis test such that rejecting the null hypothesis means detecting an alert. This allows the identification of alerts at any level within the dose region. As a result, more precise doses are determined, thereby providing a more accurate answer to the research question regarding the alert of interest. Möllenhoff et al. (2022) further developed this procedure, creating a more flexible framework that drops the assumption of monotony and includes a broader class of parametric models, thus increasing its applicability to different data structures. Key to this extension is the development of a two-step bootstrap procedure that estimates a simultaneous confidence band for the parametric model. Additionally, the procedure allows for simultaneous inference, enabling inference across the entire dose range.

This thesis builds on the concept of model-based alert identification and extends it in two directions. The first concerns the underlying research question. When analyzing time-response data, the identification of the whole time frame during which the change in response occurs might be of interest. An example for this is provided by the gene expression data generated in Ghallab et al. (2021). Genes were observed, which show a delayed deregulation, either reaching a plateau or continuing to deregulate over the study period. In such settings, the overall course of the response is of primary interest rather than a single change point. Similarly, evidence from toxicity studies suggests that transitions occur over a range of doses rather than at single dose levels (Slikker Jr. et al. 2004). Approaches such as those introduced in Kappenberg et al. (2021) and Möllenhoff et al. (2022) are limited to the identification of the onset of an effect. In contrast, this dissertation aims to shift the approach to the identification of the entire covariate region, here time frame. This is achieved by formulating the analysis in terms of the first derivative and constructing a corresponding hypothesis test. As a result, rejection of the null hypothesis means detecting an entire covariate region during which a change occurs. Thereby, the benefits of the broad applicability and interpolation between observed time points are retained while extending the analysis beyond individual points.

The second direction extends model-based alert identification to two or more continuous covariates. Incorporating more than one covariate is important for many applications, such as drug combination studies in cancer research (Mokhtari et al. 2017) or the assessment of celltoxicity (Gu et al. 2018) in pre-clinical research. Recent advancements in surface modeling for multidimensional response data (Zhou et al. 2024; Duda et al. 2022a) provide a foundation for this extension. The identification of alerts in multidimensional settings is of particular importance, as it allows for the analysis of more complex

biological mechanisms. For example, cytotoxic effects of compounds often depend jointly on exposure duration and dose Gu et al. (2018). The identification of alerts restricted to a specific exposure period therefore could result in an incomplete assessment of the toxic potential. This thesis aims to introduce an approach for model-based alert identification that accounts for the complexity in three-dimensional data settings and naturally generalizes to more dimensions. This is achieved by combining hypothesis testing based on simultaneous confidence bands with the flexible framework of GAMLSS. Depending on the specific research question, the hypothesis can be formulated in different ways. One determines the alert for one covariate while the other is fixed. The other aims to detect an alert relationship between covariates. Therefore, more complex research questions based on multidimensional data can be evaluated. In addition, the utilization of simultaneous confidence regions allows for simultaneous inference across all combinations of covariates.

This thesis is structured as follows: Section 2.1 presents the methodological background. First, the concept of time-to-event analysis is discussed with a focus on the Kaplan-Meier estimator, Cox’s PH model and the differences between the HR, OR and RR. Second, the concept of model-based alert identification is introduced. Third, the framework of GAMLSS is discussed. In Section 3, the three main methodological contributions of this thesis are summarized. The first article in Section 3.1 addresses the importance of aligning the research question with the statistical method in risk estimation for time-to-event data by the introduction of a nonparametric estimator for the RR. The second and third articles address model-based alert identification. In the second article in Section 3.2, the shift in the underlying research question from individual points to covariate regions is discussed. Section 3.3 extends the discussion from one covariate to a multidimensional setting. Section 4 concludes the thesis with a discussion of the results.

2 Preliminaries

This section outlines the theoretical foundation for the topics discussed in Section 3. Firstly, time-to-event analysis is discussed. Secondly, the concept of model-based alert identification is introduced. Finally, a basis for understanding the regression framework of GAMLSS.

2.1 Time-to-event analysis

The first thematic area discussed in this thesis concerns *time-to-event analysis*, where the primary focus is the time until occurrence of a specified event of interest. Such an event can stem from a wide range of different research topics including medicine, but also engineering or economics (Klein et al. 2003). In medical research, time until death of any cause or from a specific illness is often measured. Hence, time-to-event analysis is often alternatively called by the name survival analysis.

There are two types of missing data specific to time-to-event analysis: truncation and censoring (Klein et al. 2003). This thesis will discuss the latter in the form of *right-censoring*. Right-censoring is the most common type of censoring (Kleinbaum et al. 2012). It means that the exact event time point is unknown, but it is known that the event has yet to occur at a certain time point. There can be multiple reasons for this. Among them are the end of the study, lost to follow up or withdrawal from the study (Kleinbaum et al. 2012). Right-censoring is commonly assumed to be independent from the underlying survival distribution.

As described in Klein et al. (2003), four connected quantities can be used to analyze time-to-event phenomena: the *survival function* (S), the CDF (F), the *probability density function* (f), and the *hazard rate* (h). Let X denote the random variable describing the time until the event of interest occurs. Though X can be discrete (see Klein et al. (2003) for details), it is assumed that X is continuous in the following. The survival function, the CDF, and the probability density function of X are defined as

$$\begin{aligned} S(x) &= \mathbb{P}(X > x) = \int_x^\infty f(t)dt, \\ F(x) &= \mathbb{P}(X \leq x) = 1 - S(x), \text{ and} \\ f(x) &= -\frac{dS(x)}{dx} = \frac{dF(x)}{dx}. \end{aligned}$$

The fourth quantity is the hazard rate defined by

$$h(x) = \lim_{\Delta x \rightarrow 0} \frac{\mathbb{P}(x \leq X < x + \Delta \mid X \geq x)}{\Delta x} = \frac{f(x)}{S(x)}.$$

It is a conditional measure that can be interpreted as the instantaneous rate of experi-

encing the event conditioned on survival until x .

The distribution of X can be estimated either assuming a parametric distribution, which is discussed later in this chapter, or in a nonparametric manner. The Kaplan-Meier estimator, also Product-Limit estimator, introduced in Kaplan et al. (1958), offers a nonparametric estimator for the survival function that accounts for right-censored data. Following the introduction provided in Klein et al. (2003), let $t_1 < \dots < t_D$ denote the D distinct time points at which events occurred. Let Y_i , $i = 1, \dots, D$, denote the number of individuals at risk and d_i the number of events at t_i . If $c_{[i-1, i]}$ is the number of censored individuals between t_{i-1} and t_i , it holds $Y_i = Y_{i-1} - d_{i-1} - c_{[i-1, i]}$. The value $\frac{d_i}{Y_i}$ provides an estimate of the conditional probability that an individual who has not experienced the event just prior to t_i will experience it at t_i . The Kaplan-Meier estimator is defined by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) & \text{if } t_1 \leq t. \end{cases}$$

It is a monotonically decreasing step function. It jumps at the event time points, where the jump size depends on d_i and $c_{[i-1, i]}$. The Kaplan-Meier estimator can be considered a nonparametric maximum likelihood estimator (see Kalbfleisch et al. (2002) for details), or it can be derived from counting processes (see Klein et al. (2003) for a short introduction or Kalbfleisch et al. (2002) and Andersen et al. (1993) for more details). This allows the derivation of asymptotic theory, including asymptotic normality. In addition, *Greenwood's formula* (Greenwood 1926; Klein et al. 2003) provides an estimator for the variance of the Kaplan-Meier estimator at t

$$\widehat{\text{var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

While not specific to time-to-event analysis, the *Delta method* (Oehlert 1992) is used in Section 3.1 in conjunction with Greenwood's formula and, as such, will be introduced here. To understand the formulation, the concept of *convergence in probability* is a prerequisite (van der Vaart 1998). For this, let $(X_n)_{n \in \mathbb{N}}$ denote a sequence of real-valued random variables. It is said to converge in distribution to a random variable X if

$$\mathbb{E}f(X_n) \xrightarrow{n \rightarrow \infty} \mathbb{E}f(X)$$

holds for all bounded, continuous functions f .

The delta method as presented in van der Vaart (1998, Theorem 3.1) concerns the convergence of random variables X_n at which a differentiable function is evaluated. Precisely, let $\phi : D_\phi \subset \mathbb{R}^k \mapsto \mathbb{R}^m$ be a map defined on a subset of $\mathbb{R}^k$ and differentiable at θ . Let

X_n be random vectors taking their values in the domain of ϕ . If $r_n(X_n - \theta)$ converges in distribution to X for numbers $r_n \rightarrow \infty$, then $r_n(\phi(X_n) - \phi(\theta))$ converges in distribution to $\phi'_\theta(X)$. Here, $\phi'_\theta(X)$ denotes the Jacobian matrix of ϕ evaluated at θ and multiplied from the right with X . As a direct consequence, if $\sqrt{n}(X_n - \theta)$ converges to a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\sqrt{n}(\phi(X_n) - \phi(\theta))$ converges to $\mathcal{N}(\phi'_\theta \boldsymbol{\mu}, \phi'_\theta \boldsymbol{\Sigma} (\phi'_\theta)^T)$.

Returning to specific time-to-event analysis methods, medical research often requires the comparison between different groups, in the simplest case comparing treatment vs. control. Three measurements used to evaluate such comparisons will be discussed in Section 3.1: the HR, the RR, and the OR. In a general two groups situation, these values are defined as

$$\begin{aligned} \text{Hazard ratio: } HR(t) &= \frac{h_1(t)}{h_0(t)} \\ \text{Relative risk: } RR(t) &= \frac{F_1(t)}{F_0(t)} \\ \text{(Failure) Odds ratio: } OR(t) &= \frac{\frac{F_1(t)}{S_1(t)}}{\frac{F_0(t)}{S_0(t)}} \end{aligned}$$

where h_i , S_i , and F_i , $i = 0, 1$, denote the hazard rates, survival functions and the CDFs, respectively. The HR compares the conditional instantaneous event rates of both groups. The RR describes by what factor the risk of one group is higher than of the other. The OR compares the odds of the event of interest occurring between the groups.

An often utilized method is Cox's PH model, introduced by Cox (1972), which evaluates group differences in terms of the HR. Precisely, PH are assumed across all groups, and the joint hazard rate is formulated such that

$$h(t \mid \mathbf{X}) = h_{\text{base}}(t) \exp(\boldsymbol{\beta} \mathbf{X}). \quad (1)$$

Here, $h_{\text{base}}(t)$ describes the (unspecified) baseline hazard rate shared by all groups and $\boldsymbol{\beta} \mathbf{X}$ describes the effect of the covariates on the hazard function, similar to a linear regression model. For an in-depths discussion of regression see Section 2.3. Note, that $h_{\text{base}}(t)$ depends on time, not on the covariates, and that $\exp(\boldsymbol{\beta} \mathbf{X})$ depends on the covariates, not on time. This is inherently tied to the PH assumption, which states that the hazard ratio does not depend on time. Let, for example, $\mathbf{x}$ and $\tilde{\mathbf{x}}$ denote two expression vectors. Their hazard ratio is

$$\frac{h(t \mid \mathbf{X} = \mathbf{x})}{h(t \mid \mathbf{X} = \tilde{\mathbf{x}})} = \frac{h_{\text{base}}(t) \exp(\boldsymbol{\beta} \mathbf{x})}{h_{\text{base}}(t) \exp(\boldsymbol{\beta} \tilde{\mathbf{x}})} = \exp(\boldsymbol{\beta}(\mathbf{x} - \tilde{\mathbf{x}})).$$

However, the regression framework described by Cox's PH model can be extended to alleviate this assumption. For an introduction to time-dependent extensions of Cox's PH

model see Kleinbaum et al. (2012).

The survival function of Cox’s PH model is $S_{\text{base}}(t)^{\exp(\beta\mathbf{X})}$. Since the form of the hazard rate is assumed, but the baseline hazard rate is not specified, Cox’s PH model is considered semi-parametric (Kleinbaum et al. 2012). The estimation procedure is conducted by means of maximizing the partial likelihood as described in Klein et al. (2003). It does not necessitate to specify $h_{\text{base}}(t)$.

Alternatively, time-to-event data can be analyzed by assuming an underlying CDF. There is a multitude of possible distributions (see Klein et al. (2003) for examples), including the *Weibull* distribution, the *log-logistic* distribution and the exponentiated-uniform distribution (Kuss et al. 2021), which are discussed in detail in Section 3.1. Similar to Cox’s PH model, these distributions can be used to construct *regression models for time-to-event data* (see Klein et al. (2003) for a general introduction or Kleinbaum et al. (2012) for the specific examples of the Weibull and the log-logistic distributions). There, the shape parameter (if existent) of the chosen distribution is fixed and the scale parameter is modeled depending on the covariates. For the Weibull distribution, the log-logistic distribution, and the exponentiated-uniform distribution, fixing the shape parameter automatically results in the assumption of PH, PO, and PR respectively. In contrast to Cox’s PH model, a parametric time-to-event regression model allows for a complete characterization of S , F and f . However, this comes with a set of more restrictive assumptions that may not align with the data structure.

Time-to-event regression models are estimated by maximizing the likelihood function. Let F and f denote the CDF and the probability density function, respectively. Assuming that only events and right-censoring occur in the data, the likelihood function is defined by

$$L \propto \prod_{\text{Events}} f(x) \prod_{\substack{\text{Right-censored} \\ \text{observations}}} (1 - F(x))$$

based the reasoning regarding F and f discussed above. Note that the assumption of independent censoring is important for this to hold (Klein et al. 2003).

The analysis based on parametric time-to-event regression models or the semi-parametric Cox’s PH model is closely tied to the assumptions of proportionality listed above. These, however, also dictate by which measure the risk between the groups is evaluated. In practice, the HR, OR are often interpreted as the RR (see, e.g., Klein et al. (2003), Kalbfleisch et al. (2002), and Schwartz et al. (1999)). This is incorrect (Schwartz et al. 1999; Sutradhar et al. 2018). Figure 2 illustrated the differences between the parametric PR model based on the exponentiated-uniform distribution, the Weibull proportional hazards model and the log-logistic PO model, as exemplified by the data set discussed in the supplementary material of Section 3.1. There are clear differences in the courses

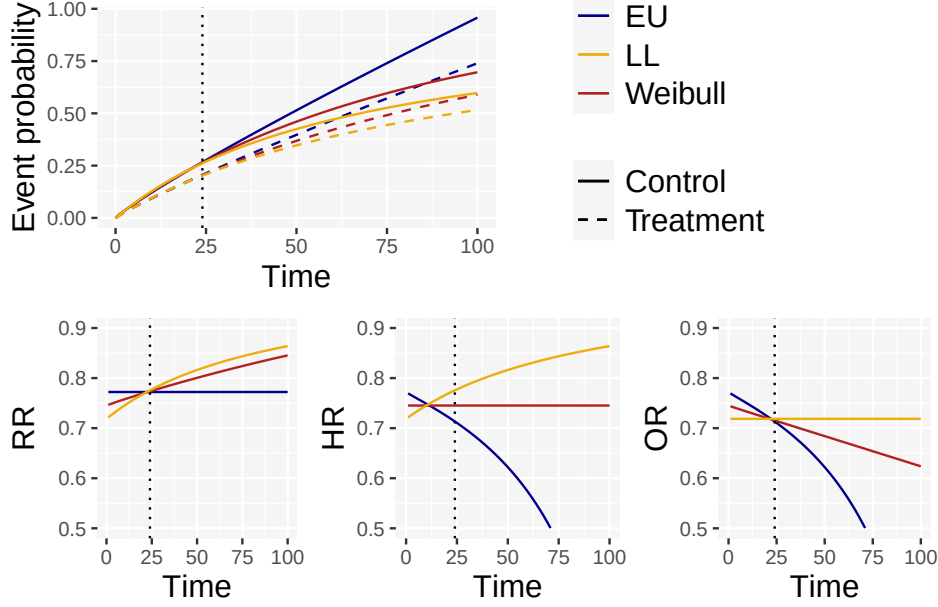


Figure 2: Comparison of the parametric PR model based on the exponentiated-uniform (EU) distribution, the Weibull PH model and the log-logistic (LL) PO model. All models are fitted to the data set discussed in the supplementary material of Section 3.1. The dotted vertical line indicates the end of study in the data set.

of the RR, HR and OR of the models, that increase with the time. Also, the constant values are not identical. This suggests that differentiation becomes more important as more information becomes available.

Section 3.1 will extensively discuss the differences between these values. A nonparametric alternative will be developed that directly estimates the RR. This alternative only assumes PR, with no additional distributional assumptions. Thus, it is applicable to a wide range of data structures.

In addition to a relative measure of risk, such as the HR, OR and RR, an absolute risk measure should be provided to allow for a holistic interpretation of the data (The Academy of Medical Sciences 2017; Zhang et al. 2018). Two of such measures are the *risk difference* (RD) and its multiplicative inverse: the *number needed to treat* (NNT). Following Altman et al. (1999), these values are defined as

$$RD(t) = F_0(t) - F_1(t) = S_1(t) - S_0(t), \text{ and}$$

$$NNT(t) = \frac{1}{RD(t)} = \frac{1}{F_0(t) - F_1(t)}$$

in a two group situation. Section 3.1 will discuss how the assumption of PR influences these values and will provide an estimator based on the proposed estimator for the RR.

2.2 Model-based identification of alerts

This section discusses the basics of model-based alert identification. To this end, the section starts with an introduction to the concept of an alert. Afterwards, the difference between *observation-based* alerts and model-based alerts is explained, as well as the difference between directly estimated alerts and those identified on the basis of statistical significance. Finally, the foundation is laid for the methods that will be introduced in Section 3.2 and 3.3. This section mainly focuses on a two-dimensional data situation. To avoid confusion, it follows the usual nomenclature of *dose* (often also *concentration*) and *response* for both alert identification and modeling. (see, e.g., Ritz et al. (2015)). However, other explanatory variables, such as *time*, as will be discussed in Section 3.1, can be evaluated equally.

2.2.1 Observation-based alerts

As formulated by Ruberg (1995a), Ruberg (1995b) and Chuang-Stein et al. (1997) and summarized in Lin et al. (2012), there are four main research questions that drive dose-response studies. These look for 1.) evidence of a drug effect, 2.) differences in responses from the response in the control group, 3.) the nature of dose-response relationships and 4.) the optimal dose. Most of these questions can be formulated as alerts or investigated through modeling (Bretz et al. 2008). The second thematic area of this thesis concerns methods to identify such alerts, specifically model-based identification. An *alert* can take many forms, but in general, is defined as the smallest dose level at which the response exceeds a prespecified value.

The process of alert identification can be broadly divided into two dimensions: observation-based alerts vs. model-based alerts, and the identification of alerts purely by estimation vs. the identification of alerts based on statistical significance. Observation-based alerts refer to alerts that identify only observed dose levels. Examples include the *LOEC* (lowest observed effect concentration) and *NOEC* (no observed effect concentration) (Calderazzo et al. 2019; Delignette-Muller et al. 2011; Murado et al. 2013), which are the lowest dose at which the mean response differs significantly from the mean response at the control level or the highest at which this difference is not significant. Hence, these are alerts concerning the second question in dose-response analysis. The estimation of the LOEC and NOEC often involves multiple comparison procedures (Bretz et al. 2008), such as multiple *t*-tests (Andersen et al. 1998), ANOVA (Warne et al. 2008) or a Dunnett test (see Hothorn (2016) for a discussion of the procedure and extensions). Both values have been criticized extensively for a variety of reasons (Murado et al. 2013; Van Der Hoeven 1997; Warne et al. 2008). One criticism, that holds for observation-based alerts in general by nature, is that the identification is limited to the dose levels used in the experiment. To address this issue, methods were introduced that identify alerts based on dose-response models (see, e.g., Calderazzo et al. (2019), Murado et al. (2013), and Andersen et al. (1998)).

2.2.2 Dose-response modeling

The first step in a model-based approach is fitting an appropriate *dose-response model*. Therefore, the concept of dose-response modeling must be introduced before model-based approaches to alert identification can be discussed. As stated in Ritz et al. (2015), dose-response models are a collection of statistical regression models with common mean structure. The selection is mainly driven by applications. The goal is to differentiate the relationship between the response variable and the explanatory variable (dose) from a random error, and describe it by fitting an appropriate regression model. Let $y_{i,j}$ denote the response variable, where $i = 0, \dots, m$ indicates the dose group and $j = 1, \dots, n_i$ denotes the observation within the dose group. In total, there are $n = \sum_{i=0}^m n_i$ observations. The dose levels are indicated by $d_i \in D \subset \mathbb{R}$ and stand in for any explanatory variable. The value d_0 indicates the control. With this, the regression model can be written as

$$y_{i,j} = f(d_i, \boldsymbol{\theta}) + \epsilon_{i,j}, \quad i = 0, \dots, m, \quad j = 1, \dots, n_i. \quad (2)$$

The expression $f(d, \boldsymbol{\theta})$, $d \in D$, describes a regression model for the mean response dependent on a parameter vector $\boldsymbol{\theta}$. The random errors are denoted by $\epsilon_{i,j}$ and are usually assumed to be independent identically normally distributed with mean 0 and standard deviation σ (see, e.g., Ritz et al. (2015) and Lin et al. (2012)). In general, estimation is conducted by least-squares methods or the maximum likelihood approach if a distribution is assumed for the errors (Fahrmeir et al. 2022; Ritz et al. 2008; Seber et al. 2003; Ritz et al. 2015). A technical discussion of assumptions and estimation procedures is deferred to the following section.

The regression function f can be derived from a variety of different regression models, including parametric nonlinear regression models (Bornkamp et al. 2009; Duda et al. 2022b), nonparametric methods (see, e.g., Kelly et al. (1990)), kernel based methods (Staniswalis et al. 1988), and semi-parametric methods (Yuan et al. 2011). This thesis will focus on the first two. *Parametric models* describe linear or nonlinear relationships between dose and response based on relatively few parameters that often allow for direct interpretation (Ritz et al. 2019; Jensen et al. 2019). In this thesis, parametric nonlinear regression models that were found by Duda et al. (2022b) to adequately describe gene expression data were specifically taken into consideration: the sigmoid emax or 4pLL model, the *beta* model the *exponential* model and the *quadratic* model. A more detailed explanation is included in Section 3.2. A visualization of these models and their parametrization are provided in Figure 3.

Spline-based models were included as a nonparametric alternative. In contrast to parametric models, parameters of nonparametric models are not fully interpretable due to their large number per model (Fahrmeir et al. 2022). There are many different types of splines-based models, the most basic of which are called basic splines or in short *B-splines*. The next paragraph closely follows the introduction presented in Fahrmeir et al.

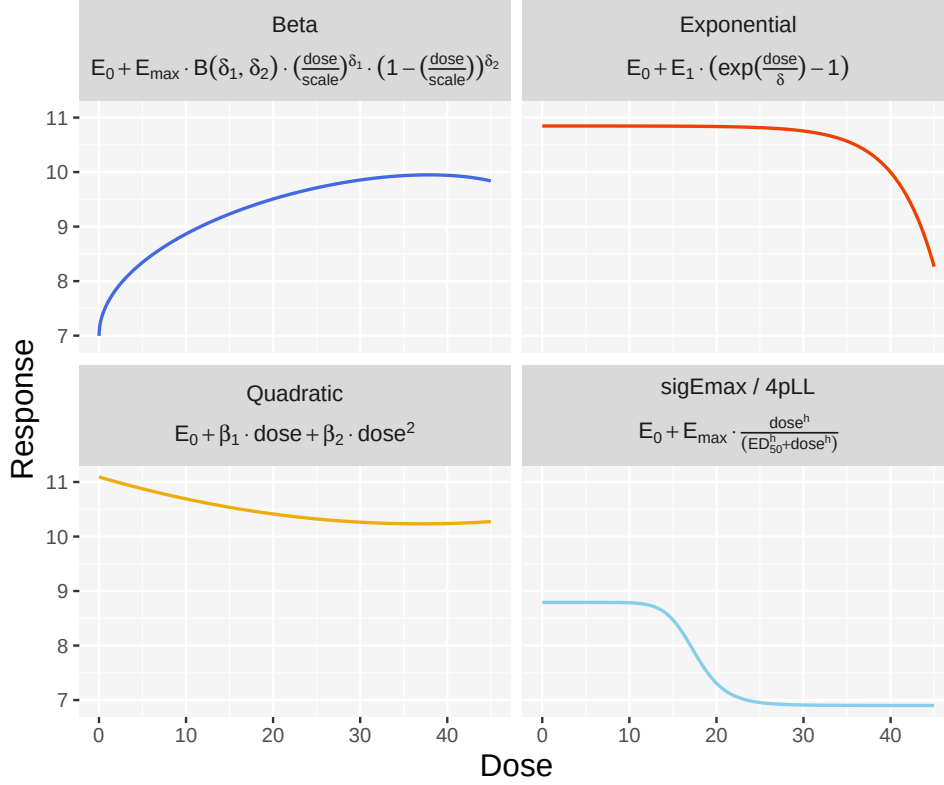


Figure 3: Visualization of a selection of common dose-response models according to Duda et al. (2022b) as discussed in the case study of Section 3.2.

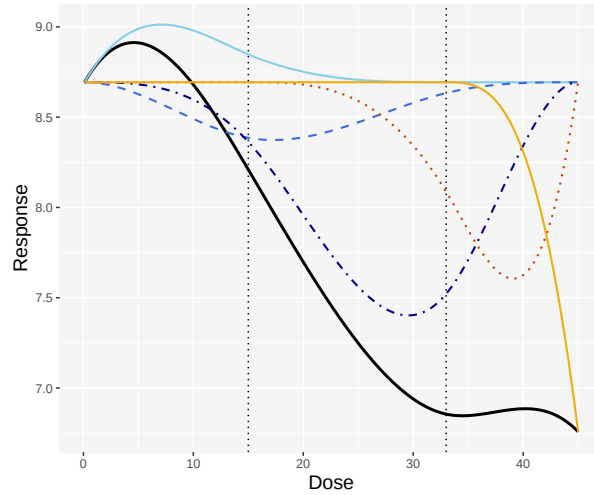


Figure 4: Example of a B-spline fit, as discussed in the case study of Section 3.2. The black line indicates the combined fit. The colored lines indicate the individual basis functions with the estimated coefficient as multiplicative factor and added intercept each for better visibility. The fit is a combination of the intercept (once) and the depicted basis functions. The knots are indicated by the vertical lines.

(2022). For the definition of splines, the covariate region, here dose range, is partitioned into intervals by *knots* $\kappa_1 < \dots < \kappa_m$. A B-spline function of degree $l \geq 0$ consists of piecewise polynomial functions of degree l in the intervals $[\kappa_j, \kappa_{j+1})$. These functions are smoothly fused at the interval boundaries, such that the overall function is $(l - 1)$ -times

continuously differentiable. For $l = 0$, the B-spline function is a step function. Given the degree and knot configurations, B-splines can be expressed as a linear combination of a set of so called basis functions. Details will be discussed in Section 3.2. Figure 4 depicts a cubic B-spline fit with two knots, as well as the corresponding basis functions. In general, cubic splines are used as a default. There are three main approaches commonly used in practice for choosing knot positions, given a set number of knots: equidistant knots, knots defined by quantiles based of the observed covariate values, and visual knot choices based on a scatter plot. Quantile-based knots are primarily placed in areas with a large number of observations, while visual knots are placed more subjectively. A higher number of knots leads to more flexible estimated functions. The number of knots can be controlled by adaptive strategies or smoothing by penalization on roughness. One penalized extension of B-Splines are so called *P-splines* (Eilers et al. 1996), which utilize penalization based on the squared second derivative of the polynomial functions. In contrast to spline based models, parametric models are in general easier to implement and take up less computation time (Jensen et al. 2019), but allow for less flexibility.

Errors in (2) are usually assumed independent identically normally distributed. This is, however, not always applicable and requires further investigation (Bretz et al. 2005). Especially in time series experiments with repeated measurements, errors are often correlated. In this case, the chosen regression model f in (2) is assumed to map a multivariate time vector to a multivariate response vector, and the errors are assumed to be independently identically multivariate normally distributed (see Seber et al. (2003) in general and Möllenhoff et al. (2018) for an application to dose-response). For further discussion on this topic, see Section 3.2. In order to reduce the number of estimated parameters, it is possible to assume specific forms of auto-correlation (Seber et al. 2003). An alternative is the inclusion of random effects (Demidenko 2013). To ensure homoscedasticity, different strategies can be applied. These include box-cox transformations of both sides (see, e.g., Calderazzo et al. (2019)), variance modeling strategies like power-of-the-mean variance models (Ritz et al. 2008; Ritz et al. 2019), robust estimation (Ritz et al. 2019), or sandwich variance estimation (Ritz et al. 2019). Some approaches investigate variance as function of the dose (West et al. 1993), e.g., assuming an exponentially increasing relationship (Jensen et al. 2019). For more complex multidimensional data structures, more flexible strategies are necessary. See Section 2.3 for further discussion.

2.2.3 Model-based alerts

Returning to the topic of alerts, model-based alerts are identified from the fitted regression model. Examples include the effective concentration, EC_p ($p\%$ of the maximal predicted effect), the *MED* (the smallest dose producing a clinically relevant response that can be declared statistically significantly different from the placebo response) (Bretz et al. 2010; Ruberg 1995a), as well as *benchmark doses* (minimal increase over the background risk) (Jensen et al. 2019). Effective concentrations and benchmark doses are parallel concepts by definition, but are usually applied to different dose regions. Benchmark doses are

more commonly used in low-dose regions (Ritz et al. 2019). Interestingly, in the context of time-to-event data as discussed in Section 2.1, the time point at which half of the events have occurred, t_{50} , is an equivalent of the EC_{50} (Ritz et al. 2019).

Hence, the first dimension of alert identification defined by observation-based vs. model-based approaches leads to a trade-off between lack of interpolation with few assumptions vs. the possibility to interpolate and a more rigid set of model assumptions that are potentially wrong (Bretz et al. 2008). Some approaches fall somewhere on the spectrum between these contrasts (see, e.g., Bretz et al. (2005) though it was later extended to a purely parametric setting with allowing for non-normal errors (Pinheiro et al. 2014) and to drug-combination settings (Zhou et al. 2024)). In addition, the classification of single alerts as model-based or observation-based is not always clear and depends on the method (see (Calderazzo et al. 2019) for a model-based approach for the identification of the LOEC or (Delignette-Muller et al. 2011) for a nonparametric approach for identifying the EC_{50}). In addition, approaches exist that try to mitigate the problems that arise from model misspecification. See Jensen et al. (2019) for an overview of model-averaging approaches for identifying benchmark doses, or Kang et al. (2000), Jensen et al. (2015) and Lin et al. (2012) for model-averaging approach for common model parameters across different models, e.g., the EC_{50} .

The second dimension of alert identification involves the estimation procedure. An alert such as the EC_{50} can be expressed as a parameter of a parametric model (Ritz et al. 2015), thus identified *estimation-based*. Alternatively, it can be estimated including a significance condition (Bretz et al. 2008), thus identified *significance-based*. Kappenberg et al. (2021) compared alerts on the four sides of these dimensions, while assuming a monotonic dose-response relationship. These are the

- *absolute lowest observed effective concentration* (ALOEC). The lowest observed concentration at which the response exceeds a pre-defined threshold (observation-based, estimation-based),
- *lowest observed effective concentration* (LOEC). The lowest observed concentration at which the response significantly exceeds a pre-defined threshold (observation-based, significance-based),
- *absolute lowest effective concentration* (ALEC). The concentration for which the inverse of the estimated 4pLL/sigmoid emax model equals a pre-defined threshold (model-based, estimation-based) and the
- *lowest effective concentration* (LEC). The lowest concentration for which the response significantly exceeds the response at the control by a pre-specified threshold. Confirmed by a hypothesis test at the α -level (model-based, significance-based).

The LOEC is estimated by multiple t -tests. For the LEC, Kappenberg et al. (2021) introduced a hypothesis-test-based method that searches the dose range for the lowest

concentration at which the following hypothesis is rejected:

$$\begin{aligned} H_0 : \forall d \in D : f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta}) &\leq \lambda && \text{for an increasing curve,} \\ H_0 : \forall d \in D : f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta}) &\geq -\lambda && \text{for an decreasing curve,} \end{aligned} \quad (3)$$

where f denotes the 4pLL or sigmoid emax model and λ denotes the pre-specified threshold. Note, that the notation was adjusted to be consistent with the notation of this section. The LEC is the lowest dose at which the hypothesis is rejected at the α -level.

The simulation study conducted by Kappenberg et al. (2021) revealed that the estimates of the model-based approaches were closer to the true alert concentration and, in general, resulted in the identification of lower doses. In addition, the significance-based approaches tended to yield slightly larger doses than the purely estimation-based approaches. The simulation study in Bretz et al. (2005) revealed similar results for more or less restrictive significance conditions when estimating the MED.

In summary, alert identification entails two dimensions with trade-offs that have to be taken into consideration based on research question and data structure:

1. Model-based alert identification vs. observation-based alert identification trading more restrictive assumptions for more precise results if these assumptions are valid.
2. Statistically significant identification of alerts vs. purely estimation based identification trading conservative results for statistical certainty to the α -level.

The estimation procedure for the LEC introduced by Kappenberg et al. (2021) assumes a monotonic relationship between dose and response and is based on fitting the 4pLL or sigmoid emax model. To remedy these limitations, Möllenhoff et al. (2022) introduced a flexible procedure that allows for a broader spectrum of models to be considered. Similar to (3), the procedure is formulated in terms of a hypothesis test

$$H_0 : \forall d \in D : |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})| \leq \lambda \text{ vs. } H_1 : \exists d \in D : |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})| > \lambda \quad (4)$$

where again λ denotes the pre-specified threshold, and f is any parametric model. Note that the notation is adjusted to be consistent with this section. Möllenhoff et al. (2022) based the test decision on the lower simultaneous confidence band. The lower simultaneous confidence band $L^\alpha(d, \hat{\boldsymbol{\theta}})$ of $|f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})|$ to the $(1 - \alpha)$ -level is defined such that

$$\mathbb{P}\{\forall d \in D : L^\alpha(d, \hat{\boldsymbol{\theta}}) \leq |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})|\} \geq 1 - \alpha.$$

Following Möllenhoff et al. (2022), H_0 in (4) is rejected if $L^\alpha(d, \hat{\boldsymbol{\theta}}) > \lambda$ and the alert is identified at $\min\{d \in D : L^\alpha(d, \hat{\boldsymbol{\theta}}) > \lambda\}$. Taking the minimum is important for a unique definition of the alert, since in contrast to Kappenberg et al. (2021), the dose-response

relationship is not assumed to be monotonic. Therefore, it is possible that $L^\alpha(d, \hat{\theta})$ exceeds λ for multiple dose levels.

The procedure is a hypothesis test at the α -level as demonstrated in Möllenhoff et al. (2022, Equation (6)). The proof proceeds as follows:

$$\begin{aligned} \mathbb{P}_{H_0}\{\forall d \in D : L^\alpha(d, \hat{\theta}) > \lambda\} &\leq \mathbb{P}_{H_0}\{\forall d \in D : L^\alpha(d, \hat{\theta}) > |f(d, \theta) - f(d_0, \theta)|\} \\ &= 1 - \mathbb{P}_{H_0}\{\forall d \in D : L^\alpha(d, \hat{\theta}) \leq |f(d, \theta) - f(d_0, \theta)|\} \\ &\leq 1 - (1 - \alpha) = \alpha. \end{aligned}$$

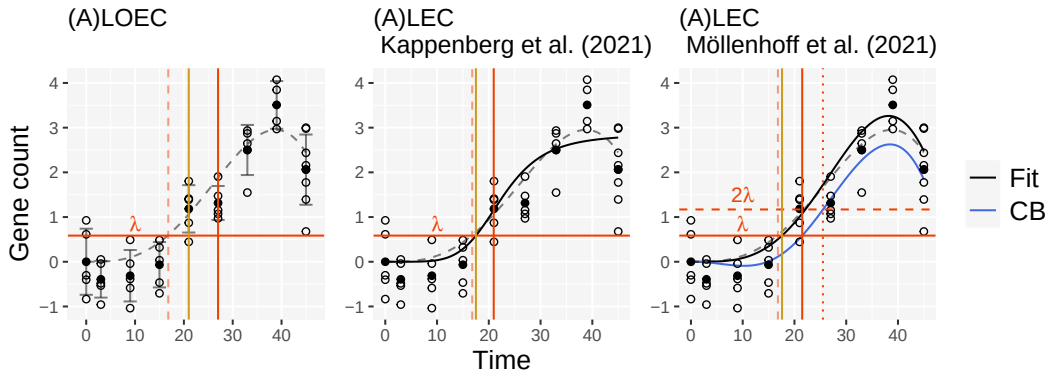


Figure 5: Comparison of observation-based alerts ALOEC/LOEC and model-based alerts ALEC/LEC. The data are taken from Scenario 5 of the simulation study discussed in Section 3.2. The truth follows the beta model and is indicated by the dashed curve. The true alert is indicated by the dashed vertical line. The rings indicate the simulated data points and the dots indicate the respective means per time point, both centered around the mean at time $d = 0$. The black curves indicate the model fits and the blue curve indicates the estimated simultaneous confidence band (CB). The red and yellow vertical lines indicate the estimated L(O)EC and AL(O)EC, respectively. For the procedure by Möllenhoff et al. (2022) a second possible threshold 2λ is investigated.

Figure 5 compares the two dimensions of alert identification using the examples of the ALOEC, LOEC, ALEC, and LEC. In addition, the results of the approaches introduced by Kappenberg et al. (2021) and Möllenhoff et al. (2022) are compared. The data are taken from Scenario 5 of the simulation study discussed in Section 3.2. The truth follows the beta model (see the supplementary material in Section 3.2) and is depicted by the dashed line. Note, that here the descriptive variable is time. The pre-specified threshold $\lambda = \log_2(1.5)$ is indicated by the horizontal red line. The true alert, which is defined as the lowest dose for which the response exceeds λ , is indicated by the dashed vertical line. The first dimension of alert identification is illustrated by comparing the results for the (A)LOEC calculated by multiple t -tests, depicted on the left, with the results for the (A)LEC calculated as described by Kappenberg et al. (2021) and Möllenhoff et al. (2022), depicted in the middle and on the right, respectively. The true alert lies between the observed time points, and thus the precision of the observation-based approach is limited by the closest. This limitation is not present in the two model-based approaches

and the estimated alerts are visibly closer to the truth. While the identified alert is close to the truth, the model misspecification present in the approach of Kappenberg et al. (2021), where the 4pLL or sigmoid emax model is assumed, misses the dip present at the end of the study. It is, however, covered by the observation-based approach. The second dimension of alert identification is illustrated by comparing the purely estimation-based AL(O)EC, denoted by a vertical yellow line, with the L(O)EC, where the threshold value is significantly exceeded, denoted by a vertical red line. The L(O)EC results are noticeably more conservative, indicating the trade-off between conservative and statistically significant results. The decision of whether or not to consider this trade-off should depend on the research question and the type of data. In some cases, conservatism might outweigh the problems posed by a conservatively identified dose level. In addition, Figure 5 depicts the benefits of a more flexible approach as introduced by Möllenhoff et al. (2022). The dip at the end of the observed time frame is recognized for by the beta model but not by the 4pLL or sigmoid emax model. However, the identified alerts are very similar. Also, the approach of Möllenhoff et al. (2022) is based on the lower simultaneous confidence band depicted in blue. This allows for simultaneous inference. To illustrate, a second threshold value 2λ is investigated, denoted by the dashed horizontal and dotted vertical lines.

The procedure introduced by Möllenhoff et al. (2022) offers several advantages beyond the expansion to non-monotonic dose-response relationships. Firstly, due to its flexible nature, it can be applied to a broad array of regression approaches, such as the parametric and non-parametric models. Secondly, while the method of Möllenhoff et al. (2022) focuses on the LEC, it allows for adjustments to identify different alerts, such as the EC_p value, by adapting the choice of λ accordingly. Finally, the approach is based on a simultaneous confidence band, thereby allowing simultaneous inference of the full curve. Therefore, inference for multiple degrees of deviation from the response at the control is possible simultaneously. In principle, the method outlined by Möllenhoff et al. (2022) is similar to approaches for simultaneous inference in benchmark dose identification. These are also often based on simultaneous confidence bands (see Jensen et al. (2019) for an overview, Deutsch et al. (2013) and Wu et al. (2006) for a selection of approaches for continuous responses and Buckley et al. (2009), Deutsch et al. (2012), and Kerns (2017) for a selection of approaches for quantal data). Many of these approaches report conservative alerts, which is unsurprising given the simultaneous nature of the inference. Thus, depending on the research question, the trade-off between simultaneous inference and conservative results is another question to be considered.

The estimation procedure for $L^\alpha(d, \hat{\theta})$ introduced by Möllenhoff et al. (2022) is based on a nested two-step bootstrap procedure. It was inspired by the bootstrap- t procedure for the estimation of confidence intervals introduced in Efron et al. (1994). A short introduction to the bootstrap is given in the following that is closely aligned to Davison et al. (1997) and Efron et al. (1994).

2.2.4 The bootstrap

The goal of a *bootstrap* is to derive reliable estimates for standard errors, confidence intervals, and other measures, if analytical calculations are complicated, approximations are not available or untrustworthy due to small sample sizes (Davison et al. 1997). As described by Efron et al. (1994) or Davison et al. (1997), the basic scenario comprises a random sample $(x_1, \dots, x_n)$ of realizations from identically distributed random variables with unknown CDF F and an estimated parameter $\hat{\phi}$. Note that both Davison et al. (1997) and Efron et al. (1994) denote this parameter θ , but here the notation was adjusted to avoid duplication. The bootstrap is utilized to infer on the probability distribution of $\hat{\phi}$. To this end, F is estimated by $\hat{F}$, which is then used to generate a bootstrap sample. From this, the parameter $\hat{\phi}$, then denoted $\hat{\phi}^*$ to differentiate from the initial estimate, is again estimated. The procedure is repeated B times and from the resulting set of $\hat{\phi}_l^*$, $l = 1, \dots, B$, the distributional measure of interest is estimated empirically. Figure 6 depicts this general concept, exemplified for the standard deviation of $\hat{\phi}$.

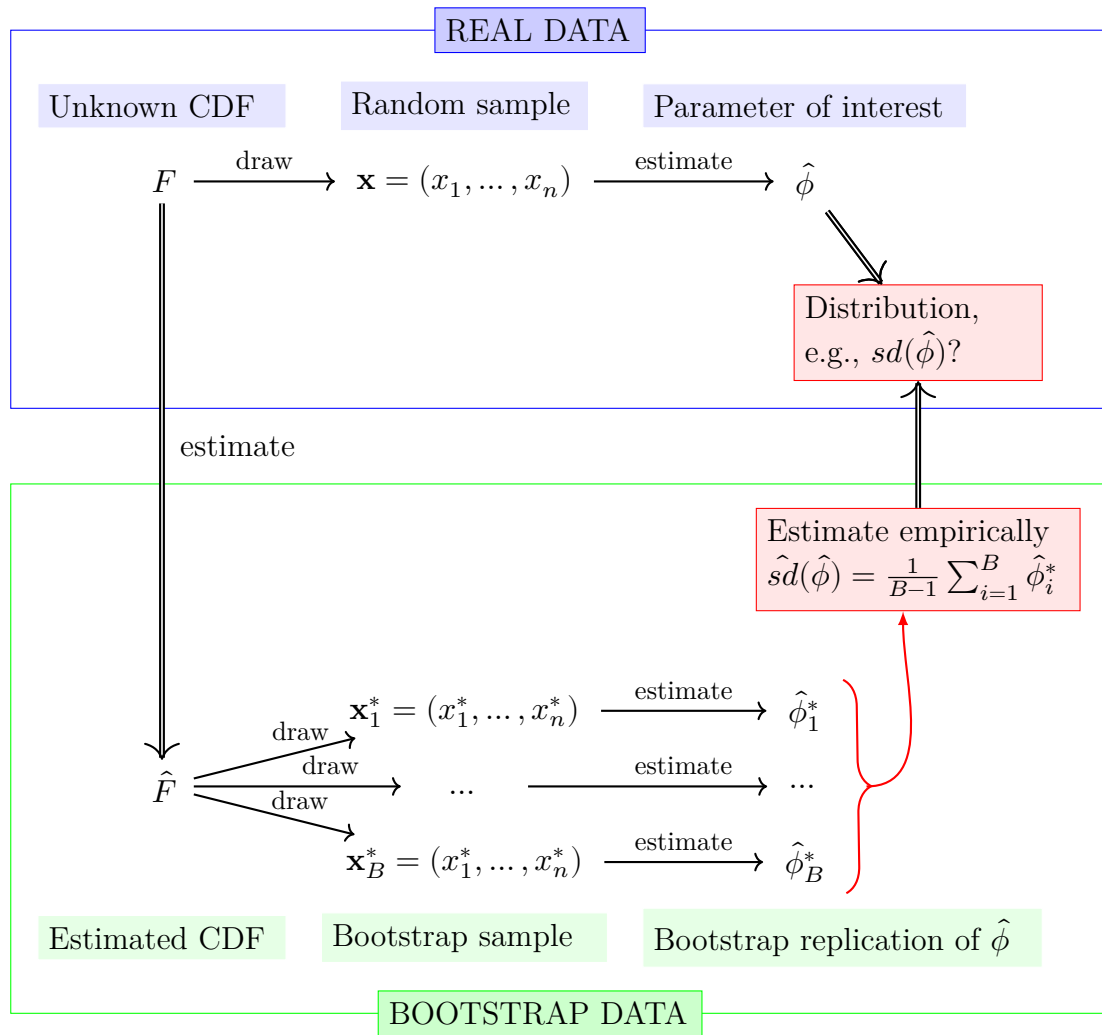


Figure 6: Visualization of the general bootstrap process, demonstrated for the standard deviation (sd) of $\hat{\phi}$. Inspired by Efron et al. (1994, Fig. 6.1 and 8.1).

The errors of the bootstrap methods are attributed to two primary sources: simulation

errors and statistical errors (Davison et al. 1997). The first results from the limited number of B bootstrap samples and asymptotically vanishes if $B \rightarrow \infty$. The second is due to the differences between F and $\hat{F}$ introduced through the estimation process. Hence, the estimation procedure is an important step (Efron et al. 1994). The influence of the assumptions made for regression models on the bootstrap will be an important topic of the discussion presented in Section 3.3. Depending on whether $\hat{F}$ is estimated nonparametrically or parametrically, the bootstrap procedure is named accordingly (Davison et al. 1997; Efron et al. 1994). The *nonparametric bootstrap* essentially involves resampling the observed realizations with replacement. The *parametric bootstrap* is based on the generation of observations from the assumed distribution. For a comparison between nonparametric and parametric approaches see, e.g., Efron et al. (1994).

Bootstrap samples used for regression analysis can be generated as follows (Efron et al. 1994; Davison et al. 1997; Carpenter et al. 2000):

- Nonparametrically, by classical resampling.
- Parametrically, assuming a model formulation as in (2) and a specific distribution. Usually normal distribution is assumed for the error terms. In this case, bootstrap samples are generated by calculating $f(d, \hat{\theta})$ from the estimated parameters and adding an error ϵ^* randomly drawn from the assumed error distribution.
- Semi-parametrically (also called nonparametrically in Efron et al. (1994)), by assuming a model as in the parametric case and calculating

$$\epsilon_{i,j} = y_{i,j} - f(d_i, \hat{\theta}).$$

Afterwards, the errors $\epsilon_{i,j}$ are resampled.

In Sections 3.2 and 3.3, the focus will lie on parametric bootstrap sample generation as it allows for a relatively simple generation of bootstrap samples in case of more complex data structures.

Many bootstrap procedures for the estimation of confidence intervals exist (see, e.g., Carpenter et al. (2000) for an overview). The *bootstrap-t* lends its form from the classical Student- t statistic or normal approximation interval (Efron et al. 1994)

$$\left[\hat{\phi} - \hat{\sigma}_{\hat{\phi}} z_{1-\alpha}, \hat{\phi} - \hat{\sigma}_{\hat{\phi}} z_{\alpha} \right], \quad (5)$$

which is derived from the assumption

$$Z = \frac{\hat{\phi} - \phi}{\hat{\sigma}_{\hat{\phi}}} \sim \mathcal{N}(0, 1).$$

In this classical case, $z_\alpha, z_{(1-\alpha)}$ are quantiles of the normal distribution or the Student- t distribution with an appropriate number of degrees of freedom. However, both types of quantiles result in symmetric intervals, thus not accounting for skewness. The bootstrap- t procedure circumvents this by bootstrapping the quantile values from the data. This is achieved by generating B bootstrap samples, computing

$$Z(\hat{\phi}_l^*) = \frac{\hat{\phi}_l^* - \hat{\phi}}{\hat{\sigma}_{\hat{\phi}_l^*}}, \quad l = 1, \dots, B \quad (6)$$

for each bootstrap sample and estimating the respective empirical quantiles from the $Z(\hat{\phi}_l^*)$. The standard deviation in both (5) and (6) can be estimated using a second-level nested bootstrap as described, e.g., in Carpenter et al. (2000).

The bootstrap- t is second order accurate (Davison et al. 1997; Efron et al. 1994). This means

$$\mathbb{P}(\phi \leq \hat{\phi} - \sigma_{\hat{\phi}} z_\alpha) = \alpha + \mathcal{O}(n^{-1}).$$

However, it is not transformation-respecting. This means that the results of the constructed confidence intervals differ depending on the scale, with some being better than others (Efron et al. 1994). It has been shown to work better in combination with variance stabilizing transformations (Efron et al. 1994) and to work well for location-parameters such as the sample mean (Efron et al. 1994). Depending on the estimated parameter, it is not necessarily range-preserving (Carpenter et al. 2000).

An alternative bootstrap estimation procedure for the calculation of confidence intervals is the percentile bootstrap (Davison et al. 1997; Efron et al. 1994). It will be utilized in Section 3.1. Here, the corresponding confidence bounds are directly estimated from the empirical quantiles $\hat{\phi}_{(\alpha)}^*$ and $\hat{\phi}_{(1-\alpha)}^*$ of the bootstrapped set $\hat{\phi}_l^*, l = 1, \dots, B$. This results in a confidence interval of the form

$$\left[\hat{\phi}_{(\alpha)}^*, \hat{\phi}_{(1-\alpha)}^* \right].$$

In contrast to the bootstrap- t it is both transformation-respecting and range-preserving (Efron et al. 1994). However, it is only first-order accurate, with a residual of order $\mathcal{O}(n^{-1/2})$ (Davison et al. 1997; Efron et al. 1994).

2.2.5 Estimation of the lower simultaneous confidence band

Returning to the approach introduced by Möllenhoff et al. (2022), the basic idea is to blow up dose-level wise bootstrap- t intervals to a confidence band encompassing the whole curve. This is achieved by the choice of a critical value c . Precisely, Möllenhoff et al.

(2022) defined

$$L^\alpha(d, \hat{\boldsymbol{\theta}}) = |f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})| - c\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|}, \quad (7)$$

where $\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|}$ denotes the estimated point-wise standard deviation of $|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|$ and c denotes the critical value. Here, c is defined such that

$$\mathbb{P}\{\forall d \in D : |f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})| - c\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|} \leq |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})|\} = 1 - \alpha.$$

This is equivalent to

$$\mathbb{P}\left\{\max_{d \in D} \frac{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})| - |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})|}{\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|}} \leq c\right\} = 1 - \alpha.$$

Estimating $L^\alpha(d, \hat{\boldsymbol{\theta}})$ necessitates the estimation of $\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|}$ and the distribution of

$$\max_{d \in D} \frac{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})| - |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})|}{\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|}}, \quad (8)$$

as c corresponds to the $(1 - \alpha)$ -quantile. In the following, denote

$$Z(d, \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})| - |f(d, \boldsymbol{\theta}) - f(d_0, \boldsymbol{\theta})|}{\hat{\sigma}_{|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|}}. \quad (9)$$

The algorithm for the estimation procedure as presented by Möllenhoff et al. (2022, Algorithm 1) is outlined in Algorithm 1 with additional details and adjusted notation for the sake of consistency.

Algorithm 1: Nested two-step bootstrap algorithm of Möllenhoff et al. (2022)

Step 1: Estimate the parameter vector $\hat{\boldsymbol{\theta}}$ for the selected model f in (2). Möllenhoff et al. (2022) utilized the *ordinary least squares* (OLS) method for the estimation, which means

$$\hat{\boldsymbol{\theta}} = \arg \min_{\mathbf{b} \in \mathcal{B}} \sum_{i=0}^m \sum_{j=1}^{n_i} \{y - f(d_i, \mathbf{b})\}^2.$$

This yields $|f(d, \hat{\boldsymbol{\theta}}) - f(d_0, \hat{\boldsymbol{\theta}})|$ for all $d \in D$. Under regularity conditions, the OLS estimator is consistent and asymptotically normally distributed (Jen-

nrich 1969). These conditions include the requirement that $f(d, \cdot)$ is a continuous function on $\mathcal{B} \subset \mathbb{R}^p$, where p is the number of parameters in the model and $\mathcal{B}$ is a compact subset of $\mathbb{R}^p$. Möllenhoff et al. (2022) further assumed the errors in (2) to be identically normally distributed, $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Hence, the OLS estimator coincides with the maximum likelihood estimator (Seber et al. 2003). Further, Möllenhoff et al. (2022) estimated σ^2 via

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=0}^m \sum_{j=1}^{n_i} \{y - f(d_i, \hat{\theta})\}^2.$$

Step 2: Generate bootstrap samples using the estimates $\hat{\theta}$ and σ^2 generated in Step 1. Precisely, draw iid. errors $\epsilon_{i,j}^* \sim \mathcal{N}(0, 1)$ and calculate

$$y_{i,j}^* = f(d_i, \hat{\theta}) + \hat{\sigma} \epsilon_{i,j}^*,$$

for $i = 0, \dots, m$ and $j = 1, \dots, n_i$. Note that this follows the idea of a parametric bootstrap procedure.

With this procedure, $l = 1, \dots, B_1$ bootstrap samples are generated. These form the basis of the first level of the bootstrap. Ultimately, the samples are used to estimate the standard deviation $\hat{\sigma}_{|f(d, \hat{\theta}) - f(d_0, \hat{\theta})|}$ and the quantile value c in equation (7). As recommended by Efron et al. (1994), B_1 should therefore be chosen relatively high. Möllenhoff et al. (2022) chose $B_1 = 500$ for their application.

Step 3: For each of the $l = 1, \dots, B_1$ bootstrap samples, estimate the parameter vector $\hat{\theta}_l^*$ and the variance $\hat{\sigma}_l^{2,*}$ as described in Step 1. Use $\hat{\theta}_l^*$ to calculate $|f(d, \hat{\theta}_l^*) - f(d_0, \hat{\theta}_l^*)|$. Then, estimate $\hat{\sigma}_{|f(d, \hat{\theta}_l^*) - f(d_0, \hat{\theta}_l^*)|}$ in equation (7) as the empirical standard deviation of $|f(d, \hat{\theta}_l^*) - f(d_0, \hat{\theta}_l^*)|$, $l = 1, \dots, B_1$.

The next goal is to estimate c in equation (7). For this, the distribution of the expression in (8) has to be estimated. Möllenhoff et al. (2022) did so by substituting $\hat{\theta}_l^*$, $l = 1, \dots, B_1$, for $\hat{\theta}$ and $\hat{\theta}$ for θ . Hence, $\hat{\sigma}_{|f(d, \hat{\theta}_l^*) - f(d_0, \hat{\theta}_l^*)|}$, $l = 1, \dots, B_1$, has to be estimated next. Möllenhoff et al. (2022) proceeded by utilizing a nested second level of bootstrapping. For each of the $l = 1, \dots, B_1$ generated first-level bootstrap samples, proceed as follows:

Step 4: Generate bootstrap samples using the estimates $\hat{\theta}_l^*$ and $\sigma_l^{2,*}$ generated in Step 3. Precisely, draw iid. errors $\epsilon_{i,j}^{**} \sim \mathcal{N}(0, 1)$ and calculate

$$y_{i,j}^{**} = f(d_i, \hat{\theta}_l^*) + \hat{\sigma}_l^* \epsilon_{i,j}^{**},$$

for $i = 0, \dots, m$ and $j = 1, \dots, n_i$. Hence, again, a parametric bootstrap sample is generated. This is repeated B_2 times. Following the recommendation of Efron et al. (1994), B_2 can be chosen relatively small, since the standard deviation is estimated. Möllenhoff et al. (2022) selected $B_2 = 25$.

Step 5: For each of the $k = 1, \dots, B_2$ second-level bootstrap samples, estimate $\hat{\theta}_k^{**}$ as described in Step 1. Use $\hat{\theta}_k^{**}$ to calculate $|f(d, \hat{\theta}_k^{**}) - f(d_0, \hat{\theta}_k^{**})|$. Then, estimate $\hat{\sigma}_{|f(d, \hat{\theta}_k^{**}) - f(d_0, \hat{\theta}_k^{**})|}$ as the empirical standard deviation of $|f(d, \hat{\theta}_k^{**}) - f(d_0, \hat{\theta}_k^{**})|$, $k = 1, \dots, B_2$.

Returning to the first level of the bootstrap:

Step 5: For each of the B_1 first-level bootstrap samples calculate

$$D_l^* = \max_{d \in D} Z(d, \hat{\theta}_l^*, \hat{\theta}) = \max_{d \in D} \frac{|f(d, \hat{\theta}_l^*) - f(d_0, \hat{\theta}_l^*)| - |f(d, \hat{\theta}) - f(d_0, \hat{\theta})|}{\hat{\sigma}_{|f(d, \hat{\theta}_l^*) - f(d_0, \hat{\theta}_l^*)|}}$$

for $l = 1, \dots, B_1$ to approximate the distribution of the expression in (8). The critical value c from (7) is estimated as the $(1 - \alpha)$ -quantile from the order statistic of D_l^* , $l = 1, \dots, B_1$.

Figure 7 depicts the estimated values throughout the process outlined in Algorithm 1. The data are taken from Scenario 5 discussed in Section 3.2. The top left depicts the original fit to the data (black) and 100 of 500 randomly selected fits to generated first-level bootstrap samples. These are used in Step 3 of Algorithm 1 to estimate $\hat{\sigma}_{|f(d, \hat{\theta}) - f(d_0, \hat{\theta})|}$, which is depicted on the top right. On the lower left, $Z(d, \hat{\theta}_l^*, \hat{\theta})$, $l = 1, \dots, B_1$, as defined in Step 6 is depicted for 100 of 500 first-level bootstrap samples. The respective maximal values D_l^* are indicated by the red dots. On the lower right, the distribution of the D_l^* for all 500 bootstrap samples is depicted in terms of a histogram. In both bottom row plots, the estimated 95%-quantile, c , is indicated by a horizontal red line.

Section 3.2 will develop a procedure that is inspired by parametric model-based alert identification with an estimation procedure that includes a significance condition. Specifically, the method will build on the groundwork laid out by Möllenhoff et al. (2022). However, the underlying research question will be shifted from identifying a single alert (dose) level to identifying entire alert regions, during which a change in response occurs. In toxicity studies, Slikker Jr. et al. (2004) showed that transitions occur over a range of doses instead of single points of departure. In addition, a covariate region of significant change is of particular interest when time is the explanatory variable, as the entire response process may be of interest. The procedure introduced in Section 3.2 will be formulated in terms of a hypothesis test based on the first derivative of a regression model. The test decision is made on the basis of a lower simultaneous confidence band for the first deriva-

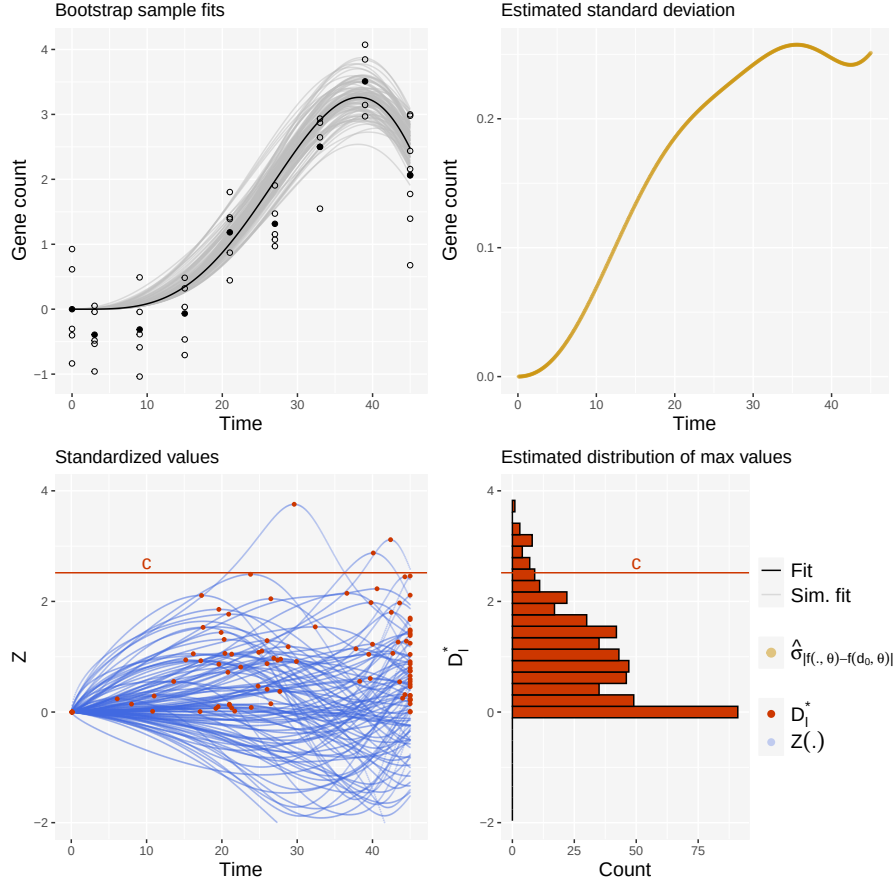


Figure 7: Visualization of the estimated values for the nested procedure introduced by Möllenhoff et al. (2022) for estimating lower simultaneous confidence bands. The data are taken from Scenario 5 of the simulation study discussed in Section 3.2. The bootstrap procedure was conducted with $B_1 = 500$ bootstrap samples for the first-level and $B_2 = 25$ bootstrap samples for the second-level. Both plots on the left depict a random selection of 100 first-level bootstrap samples for better visibility.

tive. To demonstrate applicability to different data structures, it will be investigated for parametric models as well as B-splines. The first derivative of a regression model is a natural measure for the change and has been used to define benchmark doses for continuous experimental endpoints (Sand et al. 2006).

2.2.6 Model-based alerts in a multidimensional setting

Up to this point, this section has focused on alert identification in dose-response settings. However, sometimes more data is available. In general, taking a broader approach that incorporates more complex data structures can provide a clearer view of underlying biological processes. Section 3.3 will introduce an approach to parametric model-based alert identification based on a hypothesis test in multivariate settings. It will be illustrated for two continuous covariates, specifically focusing on time-dose-response and drug-combination data. Note, however, that other types of covariate combinations are possible.

While the two dimensions of alert identification can, in principle, be transferred to multi-

dimensionality, the more complex setting raises two questions: how to adjust the model f in (2) and how to define an alert for more than one covariate. The selection of a model f depends on the analytical approach. In general, multiple fundamental approaches to the analysis can be differentiated:

- Aggregating the data for one of the covariates (e.g., time) thus assuming there is no interaction effect (Duda et al. 2022a).
- Individual analysis considering one covariate as a factor (e.g., time points) and latter combination of the results (Jiang et al. 2014; Ritz et al. 2014).
- Simultaneous fits to data grouped by one covariate (e.g., time points) assuming common features, such as common asymptotes (Ritz et al. 2014). Mixed-effects can be included (see, e.g., Ritz et al. (2019)).
- Full surface modeling again possibly including mixed effects (Ritz et al. 2014; Duda et al. 2022a).

Section 3.3 will discuss parametric surface modeling approaches for multivariate alert identification. As previously outlined, parametric models allow for the relatively simple generation of bootstrap samples, even in the case of complex data structures. Additionally, the potential for interpolating alerts between the measured covariate combinations will be examined.

As in the single-covariate case, a multitude of multivariate regression models are available. Again they can be broadly ordered into (generalized) linear regression models, parametric nonlinear regression models and nonlinear spline-based models. Ritz et al. (2014) and Serra et al. (2020) proposed the use of generalized linear models with interaction terms for surface modeling. Similarly, Deutsch et al. (2012) and Deutsch et al. (2013) utilized quadratic models with interaction terms to identify benchmark doses. Duda et al. (2022a) and Schüttler et al. (2019) proposed nonlinear regression models for time-dose response data based on sigmoidal shapes for the dose. Both approaches modeled the EC_{50} parameter depending on time. Similarly, Focke et al. (2017) modeled the time covariate as a sigmoid function and added a dose-dependent component. Zhu (2005) incorporated a nonlinear regression model and mixed-effects for repeated measurements into the identification of benchmark doses. Akkaya Hocagil et al. (2024) utilized a spline-based modeling approach with tensor products for benchmark analysis while adjusting for confounding factors by including propensity scores. For dose-combination studies specifically, it is possible to use nonlinear regression models for the individual drugs, adding an interaction term to model the drug-combination surface (see, e.g., Zhou et al. (2024)). One example of these models is further discussed in Section 3.3. Specific models also exist for pharmacokinetic/pharmacodynamic data (Lange et al. 2015). These concepts can be translated to the field of toxicology (Rozman et al. 2000).

In a multidimensional setting, an alert depends on all covariates. Therefore, rather than

having a single alert point, an alert corresponds to a relationship between the covariates. Full surface modeling allows for a full description of this relationship through the definition of *contour lines*. These are lines on the surface associated with the same level of response. For example, Schürmeyer et al. (2025) investigated a multivariate extension of the EC_p for two drugs with dose levels D_1 and D_2 and reference doses $d_{1,0}$ and $d_{2,0}$, respectively. This extension is defined by

$$MEC_p(\boldsymbol{\theta}) = \left\{ (d_1, d_2) \in D_1 \times D_2 : \frac{f(d_1, d_2, \boldsymbol{\theta}) - \min_{(d_{1,0}, d_{2,0}) \in D_1 \times D_2} f(d_{1,0}, d_{2,0}, \boldsymbol{\theta})}{R_{\max}} = \frac{p}{100} \right\},$$

where $R_{\max} = \max_{(d_1, d_2) \in D_1 \times D_2} f(d_1, d_2, \boldsymbol{\theta}) - \min_{(d_1, d_2) \in D_1 \times D_2} f(d_1, d_2, \boldsymbol{\theta})$. Note the adjusted notation. Schürmeyer et al. (2025) used the nomenclature of effective doses (ED_p and MED_p) instead of effective concentrations. This was changed here to maintain consistency and avoid confusion with the minimum effective dose. The further discussion that will be presented in Section 3.3 will follow the notation in Schürmeyer et al. (2025). In principle, rather than a single dose, a set of combinations of data points in the covariate regions is identified. Here, this set consists of different dose combinations $(d_1, d_2) \in D_1 \times D_2$. Similar example are investigated by Duda et al. (2022a), Zhu (2005), Deutsch et al. (2012) and Deutsch et al. (2013). Note that all these approaches are model-based and the latter two include significance conditions.

Figure 8 depicts the differences between the aggregated, individual and surface model fits and their corresponding single-covariate and multi-covariate EC_{50} alert values. The data are taken from a cytotoxicity data set generated by Gu et al. (2018) and discussed in the case study of Section 3.3. On the top left, the aggregated and individual fits are depicted in black and different shades of blue corresponding to the time points, respectively. The different EC_{50} values are marked by the solid (aggregated), dashed (Day 1), dot-dashed (Day 2) and dotted (Day 7) lines. On the top right, they are compared to the time-dose-alert relationship identified using the surface model. The surface model fit utilizing the model introduced by Duda et al. (2022a) is shown at the bottom. The differences between the aggregated and individual results are clearly visible, as is the curvature of the time-dependent result. These results suggest the presence of an interaction effect. In addition, both EC_{50} values identified on Day 1 and Day 7 fall outside the dose range identified by the surface modeling approach. Therefore, considering only one of the EC_{50} values might omit relevant information. These are provided by the surface modeling method.

Most multivariate models discussed in this section assume that the errors are independently and identically distributed according to a (multivariate) normal distribution. However, as has been noted on several occasions throughout this section, this assumption may not always be valid. This holds even more so for complex, three-dimensional data. Indicative of this, the dots indicating the individual measures in Figure 8 appear to scatter more strongly around the mean surface lower dose levels. This suggests a potential dependence

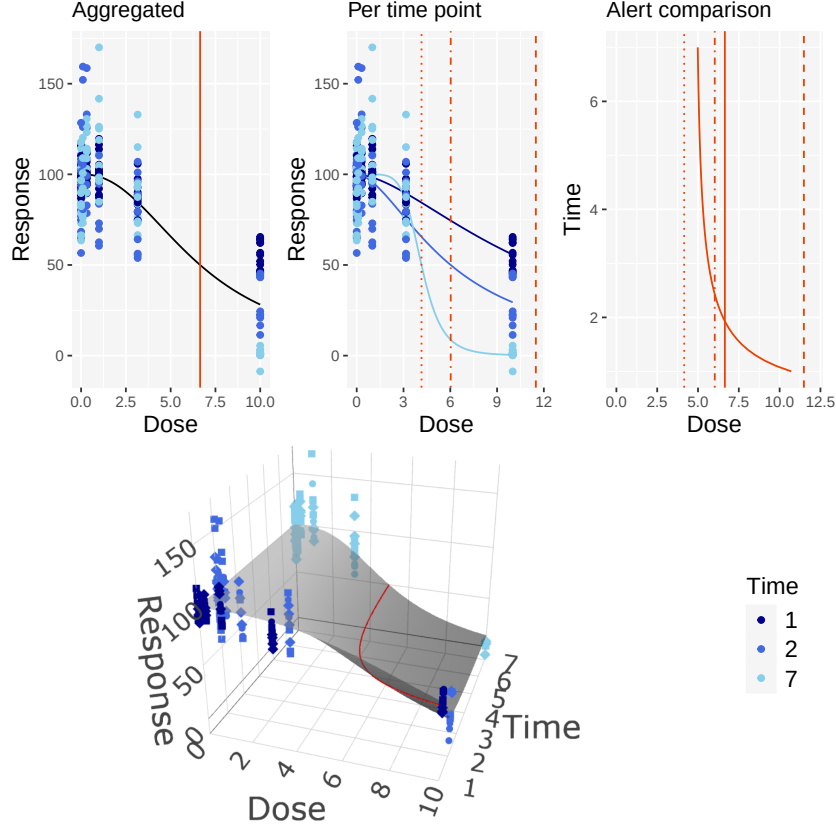


Figure 8: Visualization of different approaches for the identification of the EC_{50} in time-dose-response data inspired by the discussion of Duda et al. (2022a). The data are taken from the case study discussed in Section 3.3. The EC_{50} values are marked in red with the solid vertical line indicating the aggregated value and the dashed, dot-dashed and dotted lines indicating the values on Day 1, Day 2 and Day 7, respectively. The red curves in the upper right and lower plots denotes the time dependent EC_{50} .

of the standard deviation of the error distribution on the dose level. In the following section, the regression framework of GAMLSS is introduced, that allows modeling also for the standard deviation. It is utilized in Section 3.3 to incorporate more complex data structures in the identification of model-based alerts.

2.3 Regression modeling

This section introduces the flexible regression framework of GAMLSS (Rigby et al. 2005), which allows modeling of all parameters of the response distribution. Thus, complex relationships within the data structure can be adequately reflected. In Section 3.3, GAMLSS will be utilized as the basis for a parametric bootstrap that extends the procedure outlined in Algorithm 1. With this, a hypothesis test approach to model-based alert identification for three-dimensional data will be proposed, which naturally generalizes to multidimensional settings. However, in order to gain a better understanding of GAMLSS, the evolution of regression frameworks adapting to increasingly complex data structures is discussed first.

The underlying research question of regression analysis is to investigate the influence of

a set of covariates $(X_1, \dots, X_p)$, also known as explanatory variables, on the expected value of a response variable Y . This is achieved by estimating the systematic component $f(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$, in order to separate it from the random error ϵ (Fahrmeir et al. 2022). For the sake of simplicity, the condition on the explanatory variable is omitted from the notation in the following. Equation (2) in the previous section presented an example in which the single explanatory variable was the dose level and the response variable was continuous. In the more general setting considered in this section, let y_i , $i = 1 \dots, m$, denote m independent realizations of the response variable and $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$ the corresponding covariate vector.

2.3.1 The generalized linear model

In its simplest form, f is a linear combination of covariates, thus called a *linear regression model*. The random errors ϵ are assumed to be identically distributed with an expected value of 0 and a constant variance equal to σ^2 . In addition, the errors are often assumed to be normally distributed $\epsilon \sim \mathcal{N}(\mathbf{0}_m, \sigma^2 I_m)$. The linear model is estimated using OLS or maximum likelihood estimation, which are equivalent when a normal distribution is assumed (Fahrmeir et al. 2022). In particular, an analytical solution exists if the covariate matrix $\mathbf{X}$ has full rank (Fahrmeir et al. 2022; Wood 2017). However, the complexity of the data structures that can be represented in this way is limited. As mentioned in the previous section, extensions regarding correlated errors or heteroscedasticity exist. In these cases, OLS estimators are still unbiased, but the estimation procedure has to be adjusted to ensure that the covariance of the parameter estimates can be correctly calculated (Fahrmeir et al. 2022). A remedy procedure is the *weighted least squares* (WLS) estimation (see, e.g., Fahrmeir et al. (2022)), which extends the OLS method by including weights. The basic idea behind this is to transform the least squares problem into one with homoscedastic errors. As for traditional linear models, maximum likelihood estimation can be applied when assuming normality and coincides with WLS (Fahrmeir et al. 2022; Wood 2017). However, the weights have to be estimated as they are usually unknown. Therefore, an iterative procedure can be used that alternately estimates the weights and the parameters (*Iterative weighted least squares* (IWLS); Nelder et al. 1972; McCullagh et al. 1989; Wood 2017). Despite the different levels of complexity between the frameworks, WLS and IWLS will be a key feature of the estimation method for GAMLSS discussed in this section. However, often no closed form solution exists thus necessitating numerical approximation (see Fahrmeir et al. (2022) for a general introduction or Greene (2018) for details).

The linear model cannot accommodate many types of data structures, especially if the response is not normal. Thus, *generalized linear models* (GLM; Nelder et al. 1972) were introduced to expand the concept of linear models to other distributions. This is achieved by connecting the linear combination of covariates to the expected value of the response by means of a smooth monotonic *link function* g or the *response function* $h = g^{-1}$. Therefore,

given a response y_i , $i = 1, \dots, m$, corresponding to the covariate vector $\mathbf{x}_i$, it is

$$f(\mathbf{x}_i) = \mathbb{E}(y_i) = h(\eta(\mathbf{x}_i)) \iff g(\mathbb{E}(y_i)) = \eta(\mathbf{x}_i),$$

where $\eta(\mathbf{x}_i)$ denotes the linear predictor term (Fahrmeir et al. 2022; Wood 2017). The response is typically assumed to follow a distribution from the exponential family. Note, that the error terms can therefore depend on the mean value multiplied by a constant controlling dispersion (Wood 2017). Commonly used link functions include the logit and probit functions for binary data. The identity link function is used for normal data. IWLS or maximum likelihood with numerical approximations yield consistent and asymptotically normal estimators under regularity conditions (Fahrmeir et al. 2022; Seber et al. 2003).

2.3.2 The generalized additive model

Although the covariates can be transformed in a nonlinear manner, the GLM framework assumes that the predictor term $\eta(\mathbf{x}_i)$ is a linear combination. Thus, an important extension involves the incorporation of nonlinear relationships. The regression model class of *generalized additive models* (GAM; Hastie et al. 1986) is build around the inclusion of nonparametric nonlinear smoothing functions. Again, a link function g is utilized to connect the predictor term $\eta(\mathbf{x}_i)$ to the expected value of the corresponding response. However, the predictor is no longer assumed to be linear. In general,

$$\eta(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), \tag{10}$$

where the additive effects f_k can take a plethora of different, possibly nonlinear forms (Fahrmeir et al. 2022). These include

- basic linear effects as for the GLM,
- smooth nonlinear effects of a single continuous covariate,
- random effects. Since this thesis does not include a discussion of mixed effects models, see Demidenko (2013) for details and
- interaction effects.

In order to avoid potential issues related to identifiability, the smooth nonlinear effect terms and interaction terms including those are centered around 0. For the simple case comprising a single covariate, this means $\sum_{i=1}^m f_k(x_{i,k}) = 0$. For interaction terms, centering can be more complicated (see, e.g., Fahrmeir et al. (2022)). However, these constraints do not influence the shape (Wood 2017). The previous section introduced one example of a nonlinear, nonparametric regression function for a contentious covariate in the form of

B-splines.

In general, nonlinear smoothing functions can be expressed as a linear combination of appropriate *basis functions* B_r , $r = 1, \dots, L$, such that $f_k(x) = \sum_{r=1}^{L_k} \gamma_r B_r(x)$ (Fahrmeir et al. 2022; Wood 2017). The type of f_k determines the basis function. Section 3.2 will discuss the basis functions of B-splines in greater detail. For the observed values of the k th covariate, the linear combination can be written in matrix notation

$$\begin{pmatrix} f_k(x_{1,k}) \\ \dots \\ f_k(x_{m,k}) \end{pmatrix} = \begin{pmatrix} B_1(x_{1,k}) & \dots & B_{L_k}(x_{1,k}) \\ & \dots & \\ B_1(x_{m,k}) & \dots & B_{L_k}(x_{m,k}) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \dots \\ \gamma_{L_k} \end{pmatrix} \iff \mathbf{f}_k = \mathbf{V}_k \boldsymbol{\gamma}_k.$$

Thus, the estimation procedure for a GAM can be traced back to the estimation of (generalized) linear models (Fahrmeir et al. 2022; Stasinopoulos et al. 2017). However, to avoid overfitting, the estimation procedure includes *penalization* against roughness. The penalty is introduced through a penalty term that depends on the type of smoothing function. For example, P-splines utilize the integral of the squared second derivative of f_k . The degree of penalty is controlled by a hyperparameter $\lambda \geq 0$, such that the penalty vanishes as $\lambda \rightarrow 0$ while a subvector of $\hat{\boldsymbol{\gamma}}_k$ tends to 0 for $\lambda \rightarrow \infty$ (Stasinopoulos et al. 2024). The same general form holds true for interactions of smooth nonlinear effect (Stasinopoulos et al. 2024) and for random effects, which can be considered in terms of penalization (Fahrmeir et al. 2022). As a result, (10) can be reformulated in matrix notation (Fahrmeir et al. 2022):

$$\boldsymbol{\eta} = \mathbf{V}_1 \boldsymbol{\gamma}_1 + \dots + \mathbf{V}_q \boldsymbol{\gamma}_q + \mathbf{X} \boldsymbol{\beta}. \quad (11)$$

Here, $\mathbf{X} \boldsymbol{\beta}$ bundles the linear effects included in the model. The penalization term written as a matrix is denoted $\lambda_j \boldsymbol{\gamma}_j^t \mathbf{K}_j \boldsymbol{\gamma}_j$, with individual smoothing parameter λ_j and positive semi-definite penalization matrix $\mathbf{K}_j$ depending on the specific type of penalization. Note, that for more complex structures it is possible that λ_j is multivariate (Stasinopoulos et al. 2024). This will be omitted from the notation in the following for the sake of simplicity.

As discussed in Fahrmeir et al. (2022), the estimation procedure can be conducted based on the *penalized log-likelihood*

$$l_{\text{pen}}(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = l(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^q \lambda_j \boldsymbol{\gamma}_j^t \mathbf{K}_j \boldsymbol{\gamma}_j, \quad (12)$$

where $l(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = \sum_{i=1}^m l_i(y_i, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q, \boldsymbol{\beta})$ is the regular log-likelihood. For approximate normal response variables, a similar form can be determined for the least squares approach

$$LS_{\text{pen}}(\gamma_1, \dots, \gamma_q, \beta) = LS(\gamma_1, \dots, \gamma_q, \beta) + \sum_{j=1}^q \lambda_j \gamma_j^t \mathbf{K}_j \gamma_j, \quad (13)$$

where the least squares criterion is defined as

$$LS(\gamma_1, \dots, \gamma_q, \beta) = (\mathbf{y} - (\mathbf{V}_1 \gamma_1 + \dots + \mathbf{V}_q \gamma_q + \mathbf{X} \beta))^t (\mathbf{y} - (\mathbf{V}_1 \gamma_1 + \dots + \mathbf{V}_q \gamma_q + \mathbf{X} \beta)).$$

Assuming that the smoothing parameters λ_j are known, estimation utilizes iterative maximization of the penalized log-likelihood (Fahrmeir et al. 2001; Wood 2017; Eilers et al. 2002). Alternatively, least squares estimation is conducted by a backfitting algorithm or by means of *penalized weighted least squares* (PWLS; Fahrmeir et al. 2022). The backfitting algorithm iterates through the additive terms and estimates based on the residuals calculated from the other terms. Both the backfitting algorithm and PWLS constitute two other key features for the estimation of GAMLSS.

The determination of the smoothing parameters is integral to the estimation process. However, the discussions presented in Sections 3.2 and 3.3 neither include penalized non-parametric nonlinear effects nor random effects. Therefore, no penalization is included and the second term on the right side of equations (12) and (13) is rendered equal to 0. Thus, further discussion is not included in this thesis. For an in-depth examination on the determination of the smoothing parameters see Wood (2017).

2.3.3 Generalized additive models for location, scale, and shape

All of the regression frameworks discussed in this section focus on modeling the mean response based on the covariates. However, complex data structures may present heteroscedastic errors with complex dependence on the covariates. Figure 9 depicts two data sets from the simulation study discussed in Section 3.3. On the left, a nonlinear time-dose-response relationship assuming normal distribution with constant standard deviation is modeled. On the right, the same relationship is modeled assuming standard deviation that depends on both time and dose. The upper row depicts the simulated data points and a parametric nonlinear regression fit. The lower row depicts the corresponding residuals. This simulation is based on the real data set discussed in Section 3.3 and depicted in Figure 8. Dependence of variance on time and dose can be observed. Similar results were found in prior analysis (Duda et al. 2022a). The presented models have a limited capacity to correct for this. Depending on the response distribution, the variance can be modeled indirectly through the mean. In order to fully accommodate complex data with such dependencies, further extension of the GAM framework is necessary.

GAMLSS enable the modeling of all response distribution parameters in the manner of a GAM. For the normal distribution, this implies that both the mean and the variance can

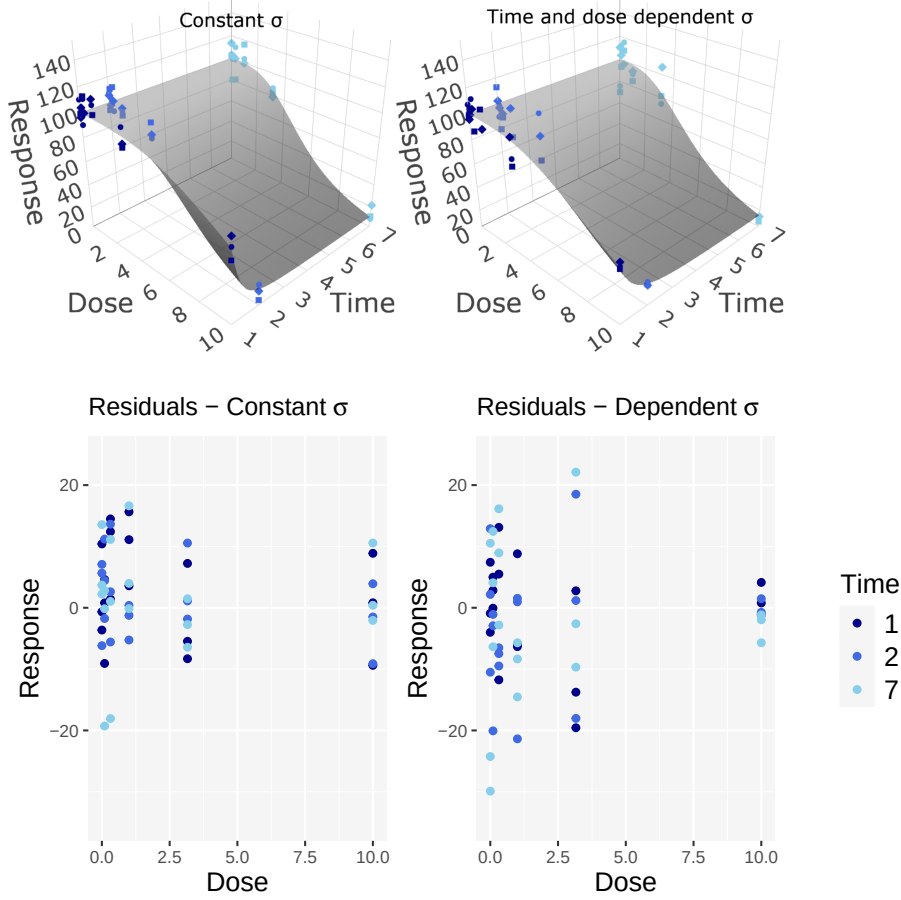


Figure 9: Visualization of the difference between a constant standard deviation across all combinations of covariates (left) with a standard deviation dependent on the covariates (right). The top row depicts the individual data points and a parametric model fit to the mean. The bottom row depicts the residuals. The data are from Scenario 1 of the simulation study discussed in Section 3.3, and are modeled such that the standard deviation is either constant or large and dependent on the covariates. The model fit utilizes the *td2pLL* model (Duda et al. 2022a).

be modeled. The GAMLSS framework is applicable for a wide range of different distributions not limited to the exponential family (Stasinopoulos et al. 2008; Rigby et al. 2019). Thus, it extends a previous approach by Rigby et al. (1996) that was only applicable to normally distributed responses. The application in this thesis will assume a normal distribution. In the interest of simplicity and generalizability, the following notation will be of normally distributed GAMLSS. The introduction mainly follows Stasinopoulos et al. (2017), while the notation was adjusted for the sake of consistency.

In general, the random variables Y_i , $i = 1, \dots, m$ are assumed to be independent and to follow a distribution $\mathcal{D}(\mu_i, \sigma_i, \nu_i, \kappa_i)$. Here, the indices suggest that the distribution parameters are dependent on the covariates. In matrix notation this can be written as $\mathbf{Y} \sim \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\kappa})$. Note, that not all distribution parameters must be modeled based on the same set of covariates and the number of distribution parameters can vary. In the

following, assume

$$\begin{aligned}
\mathbf{Y} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\
g_\mu(\boldsymbol{\mu}) &= \boldsymbol{\eta}_\mu = \mathbf{V}_{\mu,1}\boldsymbol{\gamma}_{\mu,1} + \cdots + \mathbf{V}_{\mu,q_\mu}\boldsymbol{\gamma}_{\mu,q_\mu} + \mathbf{X}_\mu\boldsymbol{\beta}_\mu \\
g_\sigma(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_\sigma = \mathbf{V}_{\sigma,1}\boldsymbol{\gamma}_{\sigma,1} + \cdots + \mathbf{V}_{\sigma,q_\sigma}\boldsymbol{\gamma}_{\sigma,q_\sigma} + \mathbf{X}_\sigma\boldsymbol{\beta}_\sigma.
\end{aligned} \tag{14}$$

Hence, both the mean and the standard deviation are modeled as described in (11) with link functions g_l , $l \in \{\mu, \sigma\}$. Again, penalization is achieved by incorporating penalization terms $\lambda_{l,j}\boldsymbol{\gamma}_{l,j}^t \mathbf{K}_{l,j}\boldsymbol{\gamma}_{l,j}$ for $l \in \{\mu, \sigma\}$ and $j = 1, \dots, q_l$ into the estimation process. For the normal distribution, g_μ is the identity function and $g_\sigma = \log(\cdot)$ is chosen to ensure values greater than 0. Rigby et al. (2005) derived the penalized log-likelihood

$$l_{\text{pen}}((\boldsymbol{\gamma}_{l,1}, \dots, \boldsymbol{\gamma}_{l,q}, \boldsymbol{\beta}_l)_{l \in \{\mu, \sigma\}}) = l((\boldsymbol{\gamma}_{l,1}, \dots, \boldsymbol{\gamma}_{l,q}, \boldsymbol{\beta}_l)_{l \in \{\mu, \sigma\}}) - \frac{1}{2} \sum_{l \in \{\mu, \sigma\}} \sum_{j=1}^{q_l} \lambda_{l,j} \boldsymbol{\gamma}_{l,j}^t \mathbf{K}_{l,j} \boldsymbol{\gamma}_{l,j}, \tag{15}$$

where $l((\boldsymbol{\gamma}_{l,1}, \dots, \boldsymbol{\gamma}_{l,q}, \boldsymbol{\beta}_l)_{l=1,2})$ denotes the regular log-likelihood of (14). In general, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are estimated for a fixed vector $\boldsymbol{\lambda}$, which is estimated in a preliminary step. However, as discussed for the GAM framework, if neither penalized smooth functions nor random effects are regarded (14) simplifies to

$$\begin{aligned}
\mathbf{Y} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\
g_\mu(\boldsymbol{\mu}) &= \boldsymbol{\eta}_\mu = \mathbf{X}_\mu\boldsymbol{\beta}_\mu \\
g_\sigma(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_\sigma = \mathbf{X}_\sigma\boldsymbol{\beta}_\sigma.
\end{aligned} \tag{16}$$

Note that B-splines are included in the linear additive term, since their definition does not incorporate penalization (Stasinopoulos et al. 2017; Fahrmeir et al. 2022)). The discussion presented in Section 3.3 will focus on this special case. Thus, estimation procedures for $\boldsymbol{\lambda}$ are not discussed here. See Stasinopoulos et al. (2017) for details.

Rigby et al. (2005) introduced two distinct estimation algorithms for maximizing (15) given known values for $\boldsymbol{\lambda}$:

- The *RS* approach based on the algorithm introduced by Rigby et al. (1996).
- The *CG* approach based on the algorithm introduced by Cole et al. (1992).

As described by Rigby et al. (2005), the CG algorithm includes the (expected or approximated) second and cross-derivatives of the log-likelihood with respect to the distribution parameters. The RS algorithm, however, does not. Since the mean and the standard deviation of a normal distribution are information orthogonal (Cox et al. 1987), the expected cross-derivatives are 0. Therefore, the RS algorithm is the preferred option in this case.

As described by Stasinopoulos et al. (2017), the RS algorithm consists of three steps: an *outer algorithm*, an *inner algorithm* and a *backfitting algorithm*. The algorithm is initialized with starting values for the distribution parameters, here μ and σ . The outer algorithm iterates through the distribution parameters and initializes the inner algorithm for each given the latest values of the others. The inner algorithm uses these latest values to conduct an IWLS approach. Precisely, the backfitting method is called to estimate the parameters $(\beta_\mu, \gamma_{\mu,1}, \dots, \gamma_{\mu,q_\mu})$, $l \in \{\mu, \sigma\}$, from (14). This is done iteratively by applying WLS or PWLS, respectively. Each step is repeated and the estimates are updated until convergence. Thus, the (penalized) maximum likelihood estimators are approximated.

The CG algorithm follows an overall similar rationale and also consists of three steps. As described by Stasinopoulos et al. (2017), the outer algorithm calculates weighted observations analogously to the inner step of the RS algorithm. Then, the inner algorithm is called, and the weighted observations are further updated. The weights used in this step include the aforementioned cross derivatives. Afterwards, the algorithm iterates through the distribution parameter and applies the backfitting method discussed above. For a proof that this procedure approximates the (penalized) maximum likelihood estimators, see Rigby et al. (2005). The CG algorithm is initially more unstable, but converges faster (Stasinopoulos et al. 2024). Therefore, an approach that starts with the RS algorithm and subsequently transitions to the CG algorithm after a series of iterations can be implemented (Stasinopoulos et al. 2017).

An important special case is described by (16). Given the specified modeling choice, the estimation process can be simplified to maximum likelihood estimation, as penalization terms are not a factor. Hence, estimators can be calculated directly by applying IWLS or maximization of the log-likelihood (Fahrmeir et al. 2022). For the normal distribution as discussed in Section 3.3 the estimation procedure benefits additionally in stability from the information orthogonality of the mean and the variance (Stasinopoulos et al. 2024). As stated by Stasinopoulos et al. (2017), the estimators for (16) are consistent with correct asymptotic standard errors under regularity conditions, given the model is correctly specified. As further elaborated by Rigby et al. (2019), the estimators are also asymptotically normally distributed. In case of model misspecification, the estimators are consistent for a distribution parameter vector of the misspecified distribution family that are closest to the true distribution measured by the Kullback-Leibler distance (Rigby et al. 2019). As a direct consequence, Rigby et al. (2019) proved that for distributions of the exponential family estimators of the mean are consistent in cases when only the dispersion parameter is misspecified. This holds for all GLMs by definition. Similarly, it holds for GAMLSS when an exponential family distribution is assumed, which includes a normal distribution. However, as argued by Stasinopoulos et al. (2017), the estimation of the mean parameters becomes inefficient. Properties of penalized maximum likelihood estimators are not as readily available (Stasinopoulos et al. 2024).

A multitude of adjustments and alternative estimation procedures have been developed

for the GAMLSS framework. Rigby et al. (2005) also investigated a Bayesian approach. See Stasinopoulos et al. (2017) and Stasinopoulos et al. (2024) for a more in-depth analysis of the Bayesian view on GAMLSS. Furthermore, Rigby et al. (2019) introduced a robust estimation procedure that combats outliers. Stasinopoulos et al. (2024) developed an estimation procedure based on machine learning that works well in high-dimensional data settings. Multivariate response variables can be considered based on multivariate distributions or a combination of marginal distributions based on copulae (Stasinopoulos et al. 2024; Kock et al. 2025).

In Sections 3.2 and 3.3 the main focus will be on regression modeling with parametric nonlinear regression models. In general, these are estimated with maximum likelihood or least squares methods (Seber et al. 2003). Under regularity conditions, the estimators are consistent and asymptotically normally distributed (Seber et al. 2003). Typically, the nonlinear regression function must be twice continuously differentiable. In Section 3.2, a new method will be introduced that extends the research question posed by alert identification from individual points to covariate regions. This is achieved by fitting a nonlinear regression model to the data and investigating the change in response by means of its first derivative.

As stated by Stasinopoulos et al. (2008), parametric nonlinear terms can be incorporated into the GAMLSS framework. Precisely, the nonlinear term is estimated in one step of the backfitting algorithm described above. Since no penalization is required for the estimation procedure, the theoretical considerations presented for GAMLSS including only non-penalized terms extends to this case. In Section 3.3, the integration of nonlinear regression models within the GAMLSS framework for the purpose of modeling time-dose-response or dose-dose-response data will be examined. The benefits of adapting to complex data structures by modeling all distribution parameters are discussed in the context of multidimensional model-based alert identification. Furthermore, an extension of the bootstrap procedure discussed in Algorithm 1 will be introduced that makes use of GAMLSS to estimate a simultaneous confidence plane for the assumed nonlinear time-dose-response or dose-dose-response relationship. Formulated as a hypothesis test, this is then utilized to identify alert relationship between the covariates based on statistical significance.

3 Articles

This section summarizes the three main methodological contributions of this thesis. Precisely, the first article in Section 3.1 discusses risk estimation in time-to-event data and the second and third article in Sections 3.2-3.3 discuss model-based alert identification for one and multiple continuous covariates, respectively.

Section 3.1 investigates the differences between the HR, OR and RR and the problems that arise from the practice of interpreting the HR and OR as the RR. To allow for a direct estimation of the latter for time-to-event data in a two-group situation, Kuss et al. (2021) proposed a parametric approach under the assumption of PR, that is a constant ratio of the CDFs. However, this method additionally assumes that the data follow an exponentiated-uniform distribution, thereby limiting its applicability. Building on the ideas of PR and the direct estimation of the RR, Section 3.1 proposes a PR approach based on the nonparametric Kaplan-Meier estimator, which requires only the PR assumption. In contrast to the approach of Kuss et al. (2021), the proposed estimator is more widely applicable and remains interpretable in most cases even when the PR assumption is violated, as it estimates an average RR over time. Additionally, an estimator for the NNT is derived, allowing for a more holistic interpretation by combining relative and absolute measures.

Section 3.2 introduces a model-based approach for alert identification with respect to a single continuous covariate, aiming to identify entire covariate regions over which the response changes. Previous works by Kappenberg et al. (2021) and Möllenhoff et al. (2022) address the limitations of traditional approaches, which only allow the detection of alerts at observed time points, by fitting parametric regression models to the data. However, these approaches merely identify the onset of an effect. In contrast, Section 3.2 proposes a method that shifts the underlying research question from the identification of individual alerts to the investigation of the overall course of the response. This is achieved by basing the analysis on the first derivative of the model rather than on the model itself. The method is summarized in terms of a hypothesis test such that only relevant changes are detected.

Section 3.3 extends the discussion on model-based alert identification to a two-dimensional setting. Inspired by Kappenberg et al. (2021) and Möllenhoff et al. (2022) a parametric approach is proposed that accommodates the complexity of multidimensional data by incorporating the flexible framework of GAMLSS into the model fitting. It generalizes naturally to higher-dimensional settings. Depending on the primary research question, the method is formulated in terms of a hypothesis test such that the rejection of the null hypothesis indicates either the detection of an alert for one of the covariates at a fixed value of the other, or the detection of an alert-relationship between the covariates. This improves interpretability by explicitly capturing interactions. Again, only relevant changes are detected.

3.1 A nonparametric proportional risk model to assess a treatment effect in time-to-event data

This article addresses the importance of aligning the research question with the statistical method in risk estimation for time-to-event data. It introduces a nonparametric estimator for the RR of two groups to experience the event of interest, based on a weighted mean of the ratio of two Kaplan-Meier estimators evaluated at different time points. Compared to existing approaches, which typically rely on the assumption of PH or PO, the proposed procedure provides an alternative for data exhibiting PR and allows for the direct interpretation of the RR. The performance of the estimator is evaluated in a simulation study and compared to established methods under various model assumptions, when RR, OR, and HR are treated interchangeably to reflect common practice. The approach's practical relevance is demonstrated through the application to data from a large randomized controlled trial investigating the effect of dapagliflozin on all-cause mortality.

Authorship:	First author
Coauthors:	Oliver Kuss, Annika Hoyer and Kathrin Möllenhoff
Contribution statement:	My contributions to this project included developing the method and implementing it in R, including programming, executing the simulation studies and case study, preparing figures and tables, and writing the first draft of the manuscript. The initial idea of a nonparametric estimator for PR came from me and was further developed with the help of Kathrin Möllenhoff. Oliver Kuss and Annika Hoyer provided the case study data set, supported the analysis and advised on the simulation scenarios. All coauthors discussed the contents and structure and revised the manuscript.
Status:	Published
Journal:	Biometrical Journal
DOI:	10.1002/bimj.202300147

RESEARCH ARTICLE

A nonparametric proportional risk model to assess a treatment effect in time-to-event data

Lucia Ameis^{1,2}  | Oliver Kuss³  | Annika Hoyer⁴  | Kathrin Möllenhoff^{1,2} 

¹Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne, Cologne, Germany

²Mathematical Institute, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

³Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany

⁴Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Bielefeld, Germany

Correspondence

Kathrin Möllenhoff, Institute of Medical Statistics and Computational Biology (IMSB), University of Cologne, Cologne, Germany.

Email: kathrin.moellenhoff@uni-koeln.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: RTG 2624, 427806116



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Time-to-event analysis often relies on prior parametric assumptions, or, if a semiparametric approach is chosen, Cox’s model. This is inherently tied to the assumption of proportional hazards, with the analysis potentially invalidated if this assumption is not fulfilled. In addition, most interpretations focus on the hazard ratio, that is often misinterpreted as the relative risk (RR), the ratio of the cumulative distribution functions. In this paper, we introduce an alternative to current methodology for assessing a treatment effect in a two-group situation, not relying on the proportional hazards assumption but assuming proportional risks. Precisely, we propose a new nonparametric model to directly estimate the RR of two groups to experience an event under the assumption that the risk ratio is constant over time. In addition to this relative measure, our model allows for calculating the number needed to treat as an absolute measure, providing the possibility of an easy and holistic interpretation of the data. We demonstrate the validity of the approach by means of a simulation study and present an application to data from a large randomized controlled trial investigating the effect of dapagliflozin on all-cause mortality.

KEYWORDS

hazard ratio, number needed to treat, risk, time-to-event analysis, treatment effect

1 | INTRODUCTION

In medical research, time-to-event data measuring the time until a specific event, for example, the time until death or the time until occurrence of a particular disease, are very common. In 2022, nearly 1000 articles including the phrase “time-to-event” were collected on PubMed (NCBI, 1996). Therefore, proper analysis of this type of data is of great interest. Well-known models include proportional hazard (PH) models like Cox’s PH model (Cox, 1972) or the Weibull PH model (Kleinbaum & Klein, 2012) and proportional odds (PO) models like the log-logistic model (Kleinbaum & Klein, 2012), resulting in a focus on the hazard ratio (HR) and the odds ratio (OR) as estimates for the treatment effect. This contradicts the preference of reporting the relative risk (RR) when analyzing data given in form of 2×2 contingency tables (Sistrom & Garvan, 2004).

The RR is characterized by its easy interpretability: If two groups, say A and B, have an RR of r for an event, then group A is r times more likely to experience the event relative to group B. Consequently, an RR of 1 means an equal risk for both groups. If r larger than 1, group A has a larger risk than group B, and vice versa if it is smaller than 1 (Rosner, 2016). The RR should always be reported in combination with an absolute risk measure like the numbers needed to treat, to classify the relative measure. Once this has been done, even complex statistical results can be communicated easily (Sistrom & Garvan, 2004).

The misinterpretation of the HR as an RR has a long tradition, starting with even basic literature using the two terms interchangeably (Kalbfleisch & Prentice, 2002; Klein & Moeschberger, 2003). However, this is incorrect: both measures indicate the same direction in regards to the treatment effect and hence have similar interpretation, but are technically not the same (Sutradhar & Austin, 2018). The HR is a conditional measure, based on instantaneous rates (Hernán, 2010), while the RR is not. Therefore, the two values should be strictly distinguished and handled with care. In addition, an increase in studies reporting nonproportional hazard rates has been noted in recent years (Royston & Parmar, 2014), putting the PH assumption into question. More precisely, the Cox model and the underlying PH assumption have been criticized recently, as the PH assumption is only rarely assessed in practice, or, the model is even used regardless of the presence of nonproportional hazards (Jachno et al., 2019). Although Cox’s model evaluates the mean HR over time in this case (Struthers & Kalbfleisch, 1986) and therefore still allows for meaningful interpretation, it nonetheless should not be confounded with the mean RR.

It is well-known that the OR approximates the RR if the event of interest has a low prevalence (Sistrom & Garvan, 2004). Otherwise, these measures do not coincide, which led to widespread misinformation in the past (Schwartz et al., 1999). In addition, just as the PH assumption, the PO assumption is sometimes not suitable. To fill in the gap of the estimation of the RR for time-to-event data, Kuss and Hoyer recently proposed a parametric proportional risk (PPR) model for a two-group situation as typically given in randomized controlled trials (RCTs) by using the exponentiated-uniform (EU) distribution (Kuss & Hoyer, 2021). This circumvents the problem of the incorrect interpretation of the HR and the OR, respectively.

Based on this idea, this paper proposes a nonparametric proportional risk (NPPR) estimator for the RR that is easy to implement, robust, and in particular independent of the underlying cumulative distribution function (CDF). For the construction the Kaplan–Meier estimator, Kaplan & Meier (1958) is used, allowing for the inclusion of right-censored data. The NPPR estimator also enables inference on a time interval, which is advantageous compared to a prior nonparametric method based on the ratio of two Kaplan–Meier estimators (Lachin, 2010) that focuses on a single time point. By means of a percentile bootstrap (Efron, 1981), we calculate confidence intervals (CIs) that allow to test for equal CDFs.

In addition to an RR measure, an absolute measure is needed to enable a holistic interpretation of the data (The Academy of Medical Sciences, 2017). Therefore, estimators for the risk difference and its inverse, the number needed to treat (NNT, Hildebrandt et al., 2009; Porta, 2016), are derived from the NPPR estimator. Especially, the NNT stands out due to its simple interpretability: If the NNT equals x_t at time point t , x_t patients have to be administered the treatment, so that one more event up to t is prevented compared to the control group (Hildebrandt et al., 2009; Porta, 2016).

The paper is structured as follows: First, we introduce the NPPR model and show how to estimate the treatment effect, given by the mean RR over time. Based on this, we present a formula for the risk difference and the NNT. Afterward, we report a small simulation study that compares the NPPR estimator to the RR estimated by applying the PPR model, the HR estimated by Cox’s PH model, and the OR estimated using a log-logistic PO model. Finally, the practical usability of the model is illustrated by its application to data from the Dapagliflozin and Prevention of Adverse Outcomes in Heart Failure (DAPA-HF) trial (McMurray et al., 2019), a large RCT investigating the effect of dapagliflozin on all-cause mortality.

2 | METHODOLOGY

2.1 | The NPPR model

We consider a situation with two groups, in terms of RCTs given by a treatment (indexed by 1) and a control (indexed by 0) group, respectively. The corresponding (unknown) CDFs describing the probability of having experienced the outcome of interest up to a specific time point, are assumed to be proportional. This implies that their ratio, the RR, is constant over time, which we denote as proportional risk (PR) assumption.

To prevent mathematical and interpretative problems, we will restrict our analysis to the time interval $T = (t_{min}, t_{max}) \subset \mathbb{R}$, whereby t_{min} denotes the larger of the first hypothetical event time points in each group and t_{max} denotes the smaller of the latest hypothetical event time points. Mathematically, this is done to (1) avoid dividing by zero and (2) avoid artificially underestimating the RR. After the first of the two CDFs reaches (or asymptotically comes close to) 1, the RR will, by definition, converge to 1. Of note, a meaningful interpretation of an RR outside this time range is hardly possible. Now, let F_1 and F_0 denote the corresponding unknown CDFs in the treatment and control group, respectively. Given the PR assumption evaluating the ratio of probabilities at any time point $t \in T$, yields a treatment effect of

$$\frac{F_1(t)}{F_0(t)} = r$$

for a constant $r > 0$. From this, we define $\beta \in \mathbb{R}$ so that the equation

$$\exp(-\beta) = \frac{F_1(t)}{F_0(t)} = r$$

holds. Solving for β we obtain

$$\beta := -\log\left(\frac{F_1(t)}{F_0(t)}\right). \quad (1)$$

Moving to the log scale allows for a more symmetric interpretation of β . More precisely, a positive β corresponds to a positive treatment effect, a negative β to a negative one, and $\beta = 0$ to no effect, respectively. The main goal is to estimate β without any parametric assumptions on the CDFs. Therefore, we estimate F_1 and F_0 by means of the nonparametric Kaplan–Meier estimators $\hat{S}_1, \hat{S}_0$, using the relationship between CDFs and their corresponding survival functions. The resulting estimated CDFs will be denoted by $\hat{F}_i = 1 - \hat{S}_i$, $i = 0, 1$. As the underlying data are possibly right-censored, we need this slightly more complicated method of estimation instead of using the empirical CDFs.

Let $\tilde{T}_1$ and $\tilde{T}_0$ be the ordered sets of observed event time points for the two groups. We set $\tilde{t}_{min} = \max(\min(\tilde{T}_1), \min(\tilde{T}_0))$ and $\tilde{t}_{max} = \min(\max(\tilde{T}_1), \max(\tilde{T}_0))$. By defining

$$\tilde{T} = (t \mid t \in \tilde{T}_1 \text{ or } t \in \tilde{T}_0, \tilde{t}_{min} \leq t \leq \tilde{t}_{max}),$$

we restrict the time interval such that the fraction of $\hat{F}_1$ and $\hat{F}_0$ can be evaluated properly and to comply with the restrictions described above. Of note, tied observations are retained in $\tilde{T}_1, \tilde{T}_0$, and $\tilde{T}$ for later inclusion. Inserting the estimated CDFs instead of the true unknown CDFs in Equation (1) for any $t, \tilde{t}_{min} \leq t \leq \tilde{t}_{max}$, yields a (time-dependent) estimator $\hat{\beta}_t$ for β . Conceivably, the method might benefit from further restricting $\tilde{T}$, for example, if the data suggest a high variability at the beginning or at the end of the study.

To arrive at a single estimate for β , we compute a weighted mean of the $\hat{\beta}_t$ where the weights are derived from their variance. Precisely, as the Kaplan–Meier estimator is asymptotically normally distributed with a variance which can be approximated by Greenwood's formula (Breslow & Crowley, 1974), we estimate the variance of $\hat{\beta}_t = -\log\left(\frac{\hat{F}_1(t)}{\hat{F}_0(t)}\right) = -\log(\hat{F}_1(t)) + \log(\hat{F}_0(t))$ for every $t \in \tilde{T}$ by applying the delta method (Oehlert, 1992) to $\log(\hat{F}_i(t))$, $i = 0, 1$. This yields

$$\hat{\sigma}_i^2(t) := \widehat{\text{var}}(\log(\hat{F}_i(t))) = \frac{1}{\hat{F}_i(t)^2} \widehat{\text{var}}(\hat{S}_i(t)), \quad i = 0, 1.$$

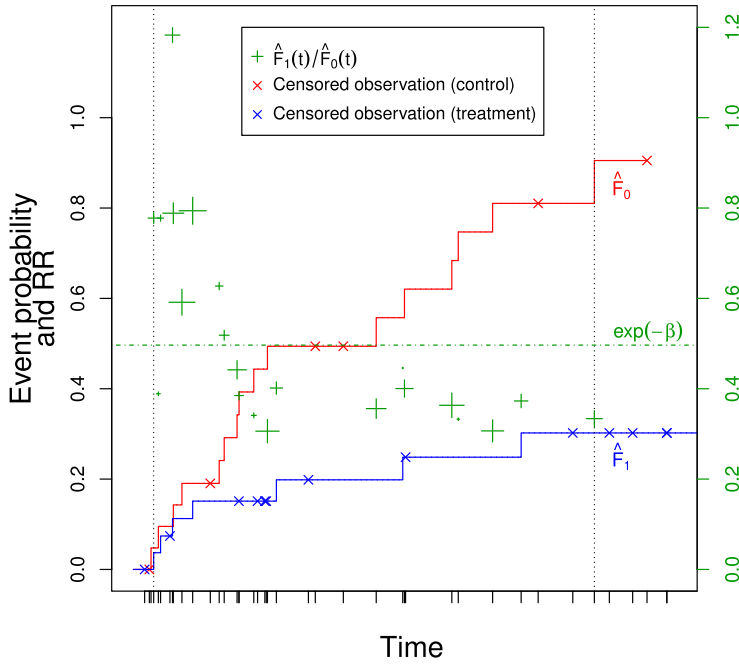


FIGURE 1 Visualization of the nonparametric proportional risk (NPPR) estimator. The size of the pluses symbolizes the weight given to the respective estimated relative risk (RR) at this time point. The larger the plus, the larger the weight. For a visualization on a logarithmic scale, we refer to the supplementary material (Figure A).

Since $\hat{F}_1(t)$ and $\hat{F}_0(t)$ are independent and consequently their covariance is equal to 0, we define the weight function by

$$\omega(t) := \widehat{\text{var}}(\hat{\beta}_t) = \frac{1}{\hat{F}_1(t)^2} \widehat{\text{var}}(\hat{S}_1(t)) + \frac{1}{\hat{F}_0(t)^2} \widehat{\text{var}}(\hat{S}_0(t))$$

for $t \in \tilde{T}$. This choice was inspired by inverse-variance weighting (Hartung et al., 2008) and the weighted least square method (Mandel, 1984). Consequently, with $W := \sum_{t \in \tilde{T}} \frac{1}{\omega(t)}$, a nonparametric estimator of β , the NPPR estimator, is defined by

$$\hat{\beta}_{NPPR} = \frac{1}{W} \sum_{t \in \tilde{T}} \frac{1}{\omega(t)} \hat{\beta}_t = -\frac{1}{W} \sum_{t \in \tilde{T}} \frac{1}{\omega(t)} \log \left(\frac{\hat{F}_1(t)}{\hat{F}_0(t)} \right). \quad (2)$$

As stated above, tied observations were retained in $\tilde{T}$. Therefore, if an event time point t' is included m times over both T_1 and T_0 , the term $\frac{1}{\omega(t')} \hat{\beta}_{t'}$ is also included m times in the sum in (2). A visualization of the situation is given in Figure 1. Also, the estimator does, in general, estimate the mean RR over time. This holds even if the PR assumption is violated. In this case, the estimator still provides a summary of the RR.

The NPPR estimator $\hat{\beta}$ is a consistent estimator for β , which is a direct consequence of its construction, the continuous mapping theorem (van der Vaart, 1998), and the fact that the Kaplan–Meier estimators are consistent for every $t \in \tilde{T}$ (Andersen et al., 1993).

2.2 | Risk difference and NNT

In order to ensure a holistic interpretation of the data, an absolute measure is needed in addition to the relative measure provided by the RR. The risk difference for the unknown true CDFs is defined by

$$RD(t) = F_0(t) - F_1(t)$$

for $t \in T$. Using the PR assumption, we can write $F_1(t) = \exp(-\beta)F_0(t)$ and substitute $F_1(t)$ in the formula. This yields

$$RD(t) = (1 - \exp(-\beta))F_0(t).$$

TABLE 1 Overview of the three different assumptions and their abbreviations.

Assumption	Abbr.	Mathematical formulation
Proportional risks	PR	It exists $r > 0$ so that $\frac{F_1(t)}{F_0(t)} = r$, with $F_i(t)$, $i = 0, 1$, as CDFs
Proportional hazards	PH	It exists $r > 0$ so that $\frac{h_1(t)}{h_0(t)} = r$, with $h_i(t)$, $i = 0, 1$, as hazard functions
Proportional (failure) odds	PO	It exists $r > 0$ so that $\frac{\frac{F_1(t)}{1-F_1(t)}}{\frac{F_0(t)}{1-F_0(t)}} = r$, with $F_i(t)$, $i = 0, 1$, as CDFs

Abbreviation: CDF, cumulative distribution function.

Now we can insert the estimated CDF $\hat{F}_0$ corresponding to the Kaplan–Meier estimator $\hat{S}_0$ and the NPPR estimator $\hat{\beta}_{NPPR}$ obtained by (2). The resulting estimator of the risk difference is defined by

$$\widehat{RD}(t) = \left(1 - \exp(-\hat{\beta}_{NPPR})\right) \hat{F}_0(t)$$

for $\tilde{t}_{min} \leq t \leq \tilde{t}_{max}$. Of note, this still does depend on the time t and given the PR assumption monotonically increases if $\exp(-\hat{\beta}_{NPPR}) < 1$ and decreases if $\exp(-\hat{\beta}_{NPPR}) > 1$, respectively.

The NNT is defined as the reciprocal of the risk difference. Therefore, an estimator is given by

$$\widehat{NNT}(t) = \frac{1}{\widehat{RD}(t)} = \frac{1}{\left(1 - \exp(-\hat{\beta}_{NPPR})\right) \hat{F}_0(t)}.$$

It is also possible to estimate the NNT based on $\hat{F}_1(t)$:

$$\widehat{NNT}_{alt1}(t) = \frac{1}{\widehat{RD}_{alt1}(t)} = \frac{1}{\left(\exp(-\hat{\beta}_{NPPR}) - 1\right) \hat{F}_1(t)}.$$

This is beneficial, for example, if the treatment group size significantly exceeds the control group size since in this case $\hat{F}_1(t)$ would be a more reliable estimator than $\hat{F}_0(t)$. Another alternative is defined using both $\hat{F}_0(t)$ and $\hat{F}_1(t)$:

$$\widehat{NNT}_{alt2}(t) = \frac{1}{\widehat{RD}_{alt2}(t)} = \frac{1}{\left(\hat{F}_0(t) - \hat{F}_1(t)\right)}.$$

However, this estimator strongly depends on the precise time point used for evaluation since it is subject to the jumps of both $\hat{F}_1(t)$ and $\hat{F}_0(t)$, while $\widehat{NNT}(t)$ and $\widehat{NNT}_{alt1}(t)$ only depend on the jumps of one of the curves. In case of a small data set, one could look at all three estimators and infer on the range of the respective CIs (calculated using the percentile bootstrap (Efron, 1981) as described below).

3 | SIMULATION STUDY

3.1 | Setting and data generation

In order to evaluate the performance of the NPPR estimator compared to an alternative method, we conducted a simulation study. As competitors we chose the PPR model for estimating the RR, the Cox's PH model estimating the HR and a log-logistic PO model estimating the OR. An overview of the different assumptions and the different models are provided in Tables 1 and 2. The PPR model is based on the EU distribution (Kuss & Hoyer, 2021), and the corresponding CDFs are given by

$$F_{EU,i}(t) = (\theta_i t)^\alpha, \quad t \leq \frac{1}{\theta_i}, \quad \text{for } i = 0, 1,$$

TABLE 2 Overview of all models. Note, that the Weibull proportional hazard (PH) model is only used for data generation and not for parameter estimation in the simulation study. More information on the relative risk (RR), the hazard ratio (HR), and the odds ratio (OR) of the different models as well as the calculation of the formula is presented in the supplementary material Section 3.0.

Model name	Abbr.	Asm.	Type	Mathematical description	$\widehat{RR}/\widehat{HR}/\widehat{OR}$
Nonparametric proportional risk model	NPPR	PR	Nonparametric	$\hat{\beta}_{NPPR} = -\frac{1}{W} \sum_{t \in \tilde{T}} \frac{1}{\omega(t)} \log \left(\frac{\hat{F}_1(t)}{\hat{F}_0(t)} \right)$	$\widehat{RR}_{NPPR} = \exp(-\hat{\beta})$
Parametric proportional risk model	PPR	PR	Parametric	$F_{EU,i}(t) = (\theta_i t)^\alpha$	$\widehat{RR}_{PPR} = \left(\frac{\hat{\theta}_1}{\hat{\theta}_0} \right)^\alpha$
Cox's proportional hazards model	Cox's PH	PH	Semiparametric	$h_{Cox}(t, X_i) = h_0(t) \exp(\beta X_i)$ baseline hazard $X_i = \begin{cases} 1 & \text{treatment} \\ 0 & \text{control} \end{cases}$	$\widehat{HR}_{Cox} = \exp(\hat{\beta})$
Log-logistic proportional odds model	LL PO	PO	Parametric	$F_{LL,i}(t) = 1 - \frac{1}{1 + \left(\frac{t}{b_i} \right)^a}$	$\widehat{OR}_{LL} = \left(\frac{\hat{b}_0}{\hat{b}_1} \right)^{\hat{a}}$
Weibull proportional hazards model	Weibull PH	PH	Parametric	$F_{Weib,i}(t) = 1 - e^{-\left(\frac{t}{\lambda_i} \right)^k}$	$\widehat{HR}_{Weibull} = \left(\frac{\hat{\lambda}_0}{\hat{\lambda}_1} \right)^{\hat{k}}$

Abbreviations: PH, proportional hazard; PO, proportional odds; PR, proportional risk.

in which the shape parameter $\alpha > 0$ is assumed to be the same for both the treatment and the control group and the scale parameters $\theta_i > 0$, $i = 0, 1$, are assumed to be group-specific. By calculating maximum-likelihood estimates using the R function *optim* for θ_i , $i = 0, 1$, and α , the estimated RR is given by

$$\widehat{RR}_{PPR} = \left(\frac{\hat{\theta}_1}{\hat{\theta}_0} \right)^\alpha.$$

Of note, the OR and HR are identical and vary over time, as can be shown by inserting the CDF into the respective definition (see supplementary material Section 3.0 for details). For the log-logistic PO model, we chose the parameterization as provided by the R package *flexsurv* (Jackson, 2016):

$$F_{LL,i}(t) = 1 - \frac{1}{1 + \left(\frac{t}{b_i} \right)^a}, \text{ for } i = 0, 1,$$

in which the shape parameter $a > 0$ and the scale parameters $b_i > 0$, $i = 0, 1$, are similarly assumed to be the same and group-specific, respectively. Again, using *optim* to calculate maximum-likelihood estimates, the estimated (failure) OR is given by

$$\widehat{OR}_{LL} = \left(\frac{\hat{b}_0}{\hat{b}_1} \right)^{\hat{a}}.$$

Outcomes of the simulation study were bias, mean squared error (MSE), and empirical coverage of the estimated RR, respectively. In addition, we evaluated numerical robustness in terms of the number of converged simulation runs. Precisely, it might happen that all events of one group have already occurred before the first event of the other group occurs. Under these circumstances, or when all participants were censored, $\tilde{T}$ might be empty. Then, the NPPR estimator is not defined. Among the simulations, these cases are few in numbers and were excluded from further analysis. Similarly, data sets yielding estimates obtained by the PPR model, Cox's PH model, and the log-logistic PO model of an absolute effect ($|\log(\widehat{RR}_{PPR})|$, $|\log(\widehat{HR}_{Cox})|$, $|\log(\widehat{OR}_{LL})|$) larger than 3 were removed. This choice was informed by the analysis of the simulated data with Cox's PH model. That is, limiting the analysis to estimated HRs smaller than 20 ($\approx \exp(3)$) was the most conservative way to exclude estimations suffering from obvious numerical issues. Of note, the NPPR estimator

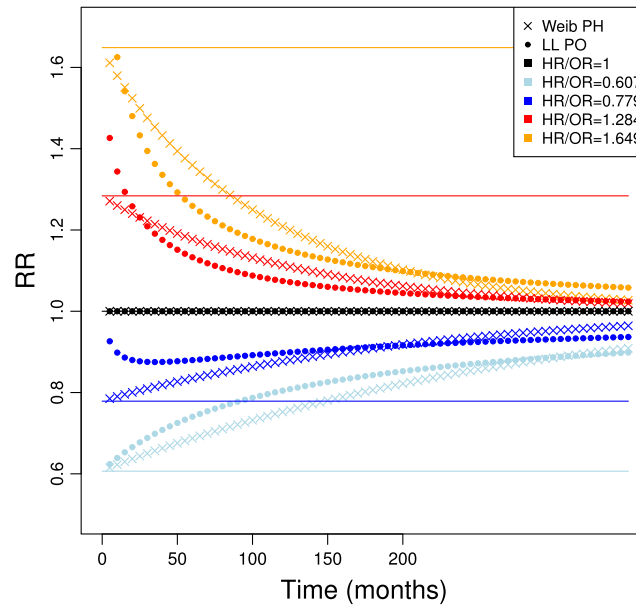


FIGURE 2 The relative risk (RR) of the Weibull proportional hazard (PH) model and the log-logistic (LL) proportional odds (PO) model over time in comparison with the constant value indicated by the straight lines. HR, hazard ratio; OR, odds ratio.

exceeded this limit in no case. The numerical robustness describes the number of estimates used for further analysis after the removal. All initial analyses have been done using R version 4.2.2. Analyses with Cox's PH model, the log-logistic PO model, and the simulation studies without censoring were added later and have been done using R version 4.3.0.

The simulation setting was inspired by the DAPA-HF trial discussed as a case study example in Section 4. Instead of focusing on the secondary outcome “death of all causes,” we based the simulations on the data set regarding the studies primary outcome “worsening heart failure or cardiovascular death.” Figure C (see supplementary material) contains an analysis of this data set analogous to Section 4. We simulated a PR, PH, and a PO situation with different true underlying effects for the RR, HR, or OR, respectively. For the generation of data satisfying the PR assumption, the PPR model was used and the log-logistic PO model for PO data. For a PH scenario, we used a Weibull PH model. The corresponding CDFs are given by

$$F_{Weib,i}(t) = \begin{cases} 1 - e^{-\left(\frac{t}{\lambda_i}\right)^k} & \text{for } t \geq 0 \text{ for } i = 0, 1. \\ 0 & \text{for } t < 0 \end{cases}$$

Here, $k > 0$ is the shape parameter, which is assumed to be equal for both groups, and λ_i , $i = 0, 1$, are the group-specific scale parameters. In all scenarios, we used the parameters obtained from the placebo group of the data set as underlying truth to simulate data for the reference group, where we chose true underlying parameters $\beta = 0, 0.5, 0.25, -0.25, -0.5$ which correspond to an RR/HR/OR equal to 1, 0.607, 0.779, 1.284, and 1.649 (see Figure 2 for a visualization of the RRs over time). All configurations used for the simulations are summarized in Table 3. For a visualization of the CDFs, we refer to Figure E in the supplementary material.

To achieve different amounts of censoring (30%, 50%, and 70%), we simulated censoring times from an appropriate uniform distribution. Also, we simulated studies without any censoring. For each of these combinations, we varied the number of study participants (50, 100, and 500). All simulation results were obtained from simulating 1000 studies. Combining the PR, PH, and PO cases resulted in 180 simulation scenarios representing a situation with no, a realistic, and a more extreme effect on both sides of the null effect. In all cases, we first drew the group assignment using inverse transformation sampling of a binomial distribution with probability $p = 0.5$. Therefore, the group sizes of the treatment and control group in the individual simulated study vary with an expected number of half the number of study participants in each group. However, the number of all patients always equals n . Afterward, corresponding to the declared underlying

TABLE 3 Parameters of the parametric proportional risk (PPR) model, the Weibull proportional hazard (PH) model, and the log-logistic proportional odds (PO) model corresponding to the true underlying effects for all different cases.

True underlying effect		Parameters								
β	RR/HR/OR	PPR			Weibull PH			log-logistic PO		
		α	θ_0	θ_1	k	λ_1	λ_0	a	b_0	b_1
0	1	0.859	0.009	0.009	0.916	88.296	88.296	0.973	70.817	70.817
0.5	0.607	0.859	0.009	0.005	0.916	152.428	88.296	0.973	70.817	118.362
0.25	0.779	0.859	0.009	0.007	0.916	116.012	88.296	0.973	70.817	91.553
−0.25	1.284	0.859	0.009	0.012	0.916	67.201	88.296	0.973	70.817	54.777
−0.5	1.649	0.859	0.009	0.016	0.916	51.146	88.296	0.973	70.817	42.370

Abbreviations: HR, hazard ratio; OR, odds ratio; RR, relative risk.

true effect, we drew survival times from the PPR model, the Weibull PH model, or the log-logistic PO model, again using inverse transformation sampling.

In case of censoring, given the censoring proportion we then drew censoring times from the uniform distribution—for the specific parameters, see the supplementary material (Table A) - for each participant. Finally, the observed time of each individual was defined as minimum of the survival and censoring time and the status adjusted accordingly. In case of no censoring, the observed time of each individual is equal to the survival time.

3.2 | Estimation

We compared the NPPR estimator with the RR estimated using the PPR model only if for the underlying data-generating model the PR assumption was satisfied. Due to numerical problems, we fitted the PPR model only if censoring was present. For the performance of the PPR model in case of PH data, we refer to Kuss & Hoyer (2021). Similarly, we fitted Cox's PH model and the log-logistic PO model only for PR data as well as for the data that fulfill the respective assumption. Of note, the Weibull PH and the log-logistic PO models violate the PR assumption. However, the NPPR estimator is in this case still applicable to estimate the mean RR over time. We used this measure as substitute for an estimator of the true underlying HR/OR and preceded with the analysis if these parameters were identical. This was done to examine the behavior of the NPPR model, if the PR assumption does not hold, which can be the case in practical applications. In correspondence to this, we interpreted the estimated HR/OR for the PR data using Cox's PH and the log-logistic PO models as mean HR/OR and used it as substitute to estimate the true RR in order to mimic the often seen practice to use the RR and the HR/OR interchangeably (Schwartz et al., 1999; Sutradhar & Austin, 2018).

The PPR and the log-logistic PO models were estimated using the R function *optim* as described above. Cox's PH model was fitted using the function *coxph* from the *survival* (Therneau, 2023; Therneau & Grambsch, 2000) package.

3.3 | Results

In the following, we will present the results of the simulation study. As the NPPR model directly estimates $\hat{\beta}_{NPPR} = -\log(\widehat{RR}_{NPPR})$, we transform the estimate $\widehat{RR}_{PPR}$ obtained by the PPR model to $-\log(\widehat{RR}_{PPR})$ for an easier comparability. We preceded accordingly with the estimated $\widehat{HR}_{Cox}$ and $\widehat{OR}_{LL}$. Tables 4–6 display the bias and the MSE for the different scenarios if censoring was present, respectively. The bias and MSE corresponding to the scenarios without censoring are displayed in Table B (see supplementary material). For details on the coverage and the numerical robustness, we refer to the supplementary material (Tables C–J).

For a quick introduction, Figure 3 displays a comparison between the estimated values for a censoring rate of 50% and 500 participants if the PR assumption is fulfilled. If the true underlying effect β equals 0, the NPPR estimator has a larger interquartile range (IQR) than the PPR model with equal median. In all remaining cases, the median is closer to the true value, while the IQR is still a bit larger. We observe that the NPPR estimator produces fewer outliers compared to the PPR model. The number of outliers and the IQR of Cox's PH model and the log-logistic PO model are similar to those of the NPPR model. As expected, the median is larger in case of $\beta \neq 0$. Similar findings are achieved for situations with a censoring rate of 30% and only in case of a high censoring rate, that is 70%, the number of outliers is comparable (see

TABLE 4 Bias and mean squared error (MSE) of $\hat{\beta}_{NPPR}$ (NPPR, nonparametric proportional risk), $-\log(\widehat{RR}_{PPR})$ (PPR, parametric proportional risk), $-\log(\widehat{HR}_{Cox})$ (Cox's PH model), and $-\log(\widehat{OR}_{LL})$ (log-logistic PO model) if the PPR model is the true underlying model.

Effect	Censoring (%)	Participants	Bias				MSE			
			NPPR	PPR	Cox	LL	NPPR	PPR	Cox	LL
0.00	30	500	0.000	−0.006	0.000	0.000	0.011	0.140	0.013	0.026
0.00	30	100	−0.002	−0.017	−0.001	−0.006	0.052	0.365	0.065	0.132
0.00	30	50	0.010	0.021	0.007	0.010	0.092	0.436	0.139	0.250
0.00	50	500	0.005	0.001	0.002	0.006	0.016	0.005	0.018	0.030
0.00	50	100	0.004	0.007	0.008	0.011	0.082	0.047	0.093	0.162
0.00	50	50	−0.011	0.000	−0.009	−0.015	0.153	0.111	0.192	0.316
0.00	70	500	−0.003	−0.002	−0.002	−0.003	0.027	0.016	0.028	0.040
0.00	70	100	−0.011	−0.005	−0.012	−0.012	0.145	0.084	0.146	0.211
0.00	70	50	0.016	0.017	0.025	0.027	0.280	0.205	0.324	0.463
0.50	30	500	0.002	0.196	0.495	0.426	0.013	0.144	0.259	0.210
0.50	30	100	0.009	0.314	0.490	0.444	0.066	0.530	0.314	0.337
0.50	30	50	0.007	0.349	0.494	0.476	0.129	0.583	0.397	0.521
0.50	50	500	−0.004	0.070	0.367	0.378	0.015	0.020	0.151	0.171
0.50	50	100	0.003	0.193	0.359	0.399	0.084	0.203	0.222	0.316
0.50	50	50	−0.012	0.242	0.347	0.414	0.168	0.291	0.328	0.520
0.50	70	500	−0.003	0.000	0.138	0.221	0.031	0.019	0.050	0.093
0.50	70	100	−0.014	0.011	0.141	0.234	0.169	0.101	0.186	0.293
0.50	70	50	−0.033	0.037	0.162	0.269	0.288	0.237	0.342	0.529
0.25	30	500	0.003	0.107	0.305	0.217	0.011	0.118	0.105	0.071
0.25	30	100	0.017	0.160	0.309	0.248	0.051	0.358	0.158	0.183
0.25	30	50	0.002	0.212	0.293	0.234	0.118	0.422	0.229	0.356
0.25	50	500	0.004	0.046	0.209	0.206	0.014	0.010	0.059	0.069
0.25	50	100	0.009	0.104	0.204	0.221	0.071	0.092	0.129	0.196
0.25	50	50	−0.014	0.127	0.175	0.197	0.157	0.193	0.232	0.361
0.25	70	500	−0.002	−0.004	0.069	0.113	0.022	0.014	0.028	0.045
0.25	70	100	0.004	0.005	0.079	0.126	0.130	0.079	0.142	0.211
0.25	70	50	0.000	0.033	0.104	0.166	0.277	0.195	0.322	0.479
−0.25	30	500	−0.001	−0.089	−0.304	−0.215	0.011	0.144	0.104	0.070
−0.25	30	100	0.000	−0.189	−0.290	−0.222	0.056	0.365	0.145	0.181
−0.25	30	50	−0.014	−0.175	−0.298	−0.252	0.116	0.446	0.227	0.362
−0.25	50	500	0.003	−0.027	−0.183	−0.188	0.015	0.008	0.050	0.064
−0.25	50	100	0.005	−0.071	−0.166	−0.192	0.070	0.079	0.115	0.179
−0.25	50	50	0.012	−0.117	−0.162	−0.194	0.148	0.157	0.207	0.348
−0.25	70	500	−0.007	−0.003	−0.075	−0.120	0.025	0.016	0.032	0.052
−0.25	70	100	−0.025	−0.027	−0.101	−0.156	0.154	0.096	0.167	0.250
−0.25	70	50	0.001	−0.028	−0.103	−0.154	0.310	0.215	0.348	0.493
−0.50	30	500	0.003	−0.219	−0.485	−0.419	0.012	0.193	0.248	0.200
−0.50	30	100	−0.014	−0.306	−0.487	−0.454	0.061	0.467	0.307	0.334
−0.50	30	50	0.006	−0.300	−0.463	−0.448	0.119	0.600	0.358	0.478
−0.50	50	500	−0.004	−0.102	−0.377	−0.390	0.016	0.052	0.159	0.180
−0.50	50	100	0.007	−0.200	−0.353	−0.388	0.080	0.171	0.208	0.302
−0.50	50	50	0.009	−0.232	−0.359	−0.419	0.160	0.244	0.326	0.501
−0.50	70	500	−0.010	−0.008	−0.140	−0.225	0.029	0.017	0.047	0.089
−0.50	70	100	−0.005	−0.027	−0.149	−0.244	0.157	0.111	0.182	0.280
−0.50	70	50	0.053	−0.031	−0.149	−0.250	0.311	0.261	0.373	0.546

Abbreviations: LL, log-logistic; PH, proportional hazard; PO, proportional odds.

TABLE 5 Bias and mean squared error (MSE) of $\hat{\beta}_{NPPR}$ (NPPR, nonparametric proportional risk) and $-\log(\widehat{HR}_{Cox})$ (Cox's proportional hazard [PH]) if the Weibull PH model is the true underlying model.

Effect	Censoring (%)	Participants	Bias		MSE	
			NPPR	Cox	NPPR	Cox
0.00	30	500	−0.004	−0.005	0.010	0.012
0.00	30	100	−0.007	−0.001	0.049	0.063
0.00	30	50	−0.002	−0.006	0.108	0.138
0.00	50	500	−0.001	−0.004	0.016	0.018
0.00	50	100	−0.005	−0.003	0.076	0.086
0.00	50	50	−0.024	−0.020	0.157	0.179
0.00	70	500	−0.001	0.004	0.028	0.027
0.00	70	100	−0.015	−0.020	0.141	0.151
0.00	70	50	0.009	0.007	0.302	0.344
0.50	30	500	−0.128	0.002	0.027	0.012
0.50	30	100	−0.128	0.015	0.071	0.063
0.50	30	50	−0.126	0.012	0.116	0.128
0.50	50	500	−0.091	−0.001	0.026	0.018
0.50	50	100	−0.082	0.012	0.098	0.106
0.50	50	50	−0.073	0.034	0.189	0.217
0.50	70	500	−0.061	−0.001	0.036	0.032
0.50	70	100	−0.051	0.015	0.167	0.162
0.50	70	50	−0.076	0.061	0.343	0.397
0.25	30	500	−0.066	0.003	0.014	0.011
0.25	30	100	−0.066	−0.005	0.051	0.058
0.25	30	50	−0.062	0.018	0.099	0.124
0.25	50	500	−0.051	0.000	0.017	0.016
0.25	50	100	−0.047	0.000	0.075	0.086
0.25	50	50	−0.041	0.006	0.174	0.201
0.25	70	500	−0.033	0.002	0.027	0.027
0.25	70	100	−0.028	0.004	0.146	0.149
0.25	70	50	−0.022	0.035	0.309	0.358
−0.25	30	500	0.056	−0.010	0.014	0.012
−0.25	30	100	0.073	−0.003	0.053	0.064
−0.25	30	50	0.069	−0.008	0.105	0.128
−0.25	50	500	0.058	0.009	0.018	0.016
−0.25	50	100	0.050	−0.002	0.087	0.095
−0.25	50	50	0.049	−0.010	0.149	0.184
−0.25	70	500	0.030	0.002	0.032	0.031
−0.25	70	100	0.008	−0.024	0.152	0.161
−0.25	70	50	0.056	−0.016	0.321	0.344
−0.50	30	500	0.120	−0.012	0.024	0.012
−0.50	30	100	0.138	0.000	0.071	0.065
−0.50	30	50	0.119	−0.023	0.120	0.140
−0.50	50	500	0.101	−0.001	0.024	0.016
−0.50	50	100	0.096	−0.015	0.085	0.090
−0.50	50	50	0.120	−0.006	0.165	0.183
−0.50	70	500	0.060	−0.002	0.033	0.028
−0.50	70	100	0.045	−0.036	0.148	0.151
−0.50	70	50	0.103	−0.015	0.350	0.348

TABLE 6 Bias and mean squared error (MSE) of $\hat{\beta}_{NPPR}$ (NPPR, nonparametric proportional risk) and $-\log(\widehat{OR}_{LL})$ (log-logistic PO, proportional odds) if the log-logistic PO model is the true underlying model.

Effect	Censoring (%)	Participants	Bias		MSE	
			NPPR	LL	NPPR	LL
0.00	30	500	−0.002	−0.003	0.009	0.026
0.00	30	100	0.008	0.009	0.043	0.124
0.00	30	50	0.010	0.020	0.096	0.294
0.00	50	500	−0.005	−0.005	0.014	0.028
0.00	50	100	0.000	0.001	0.077	0.160
0.00	50	50	0.007	0.016	0.145	0.324
0.00	70	500	0.008	0.011	0.028	0.039
0.00	70	100	−0.013	−0.014	0.160	0.236
0.00	70	50	−0.013	−0.024	0.331	0.517
0.50	30	500	−0.206	0.009	0.052	0.027
0.50	30	100	−0.211	0.011	0.089	0.125
0.50	30	50	−0.211	0.018	0.149	0.313
0.50	50	500	−0.163	0.008	0.041	0.030
0.50	50	100	−0.164	0.013	0.100	0.160
0.50	50	50	−0.191	−0.009	0.204	0.364
0.50	70	500	−0.107	−0.002	0.039	0.038
0.50	70	100	−0.119	0.004	0.180	0.229
0.50	70	50	−0.108	0.078	0.363	0.548
0.25	30	500	−0.105	0.001	0.020	0.025
0.25	30	100	−0.111	−0.003	0.062	0.139
0.25	30	50	−0.118	−0.007	0.102	0.262
0.25	50	500	−0.082	0.001	0.021	0.031
0.25	50	100	−0.069	0.022	0.074	0.149
0.25	50	50	−0.080	0.007	0.165	0.332
0.25	70	500	−0.059	0.000	0.030	0.040
0.25	70	100	−0.046	0.025	0.146	0.210
0.25	70	50	−0.073	−0.003	0.293	0.442
−0.25	30	500	0.099	−0.010	0.020	0.026
−0.25	30	100	0.089	−0.025	0.055	0.128
−0.25	30	50	0.115	0.012	0.110	0.296
−0.25	50	500	0.090	0.012	0.023	0.030
−0.25	50	100	0.091	0.003	0.075	0.136
−0.25	50	50	0.065	−0.036	0.163	0.327
−0.25	70	500	0.056	0.001	0.030	0.039
−0.25	70	100	0.061	0.010	0.153	0.207
−0.25	70	50	0.055	−0.035	0.313	0.515
−0.50	30	500	0.208	−0.007	0.053	0.026
−0.50	30	100	0.223	0.010	0.100	0.142
−0.50	30	50	0.219	−0.010	0.154	0.327
−0.50	50	500	0.164	−0.001	0.045	0.036
−0.50	50	100	0.158	−0.016	0.103	0.159
−0.50	50	50	0.166	−0.028	0.189	0.338
−0.50	70	500	0.108	−0.011	0.039	0.040
−0.50	70	100	0.105	−0.026	0.167	0.233
−0.50	70	50	0.133	−0.028	0.319	0.481

Abbreviation: LL, log-logistic.

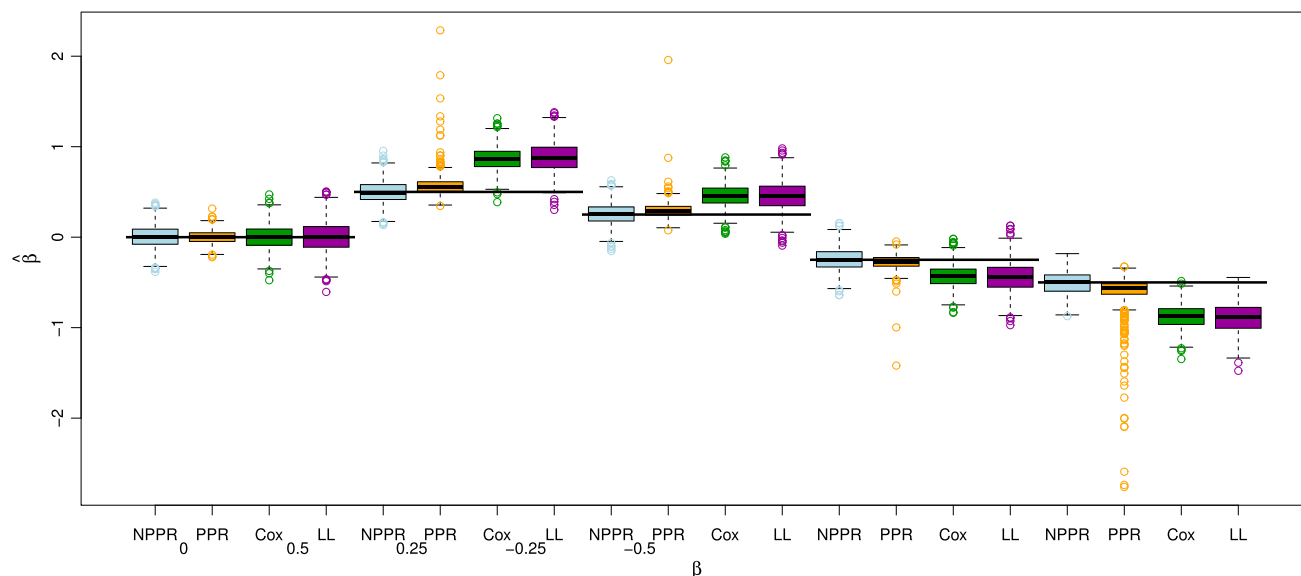


FIGURE 3 Comparison of boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR, nonparametric proportional risk), $-\log(\hat{RR}_{PPR})$ (PPR, parametric proportional risk), $-\log(\hat{HR}_{Cox})$ (Cox), and $-\log(\hat{OR}_{LL})$ (LL, log-logistic) for different choices of β if the PPR model is the underlying model, each in case of 500 participants and 50% censoring rate. Horizontal bars indicate the true underlying effect. Boxplots for 30% and 70% censoring rates and for the proportional hazard/proportional odds (PH/PO) cases are displayed in the supplementary material (Figures H–M).

supplementary material Figures F–G). For a comparison of different numbers of participants in case of a moderate effect of $\beta = 0.25$, under no censoring and assuming that all respective assumptions of the models are fulfilled, see Figure N of the supplementary material.

3.3.1 | Bias

Tables 4–6 and B (see supplementary material) depict the bias. In general, the bias is larger if the underlying model assumption is violated. In this case, a larger true effect results in a larger bias for all models.

It turns out that if the censoring rate is 50% or smaller the NPPR estimator consistently outperforms the PPR model for all underlying effects except of $\beta = 0$. This also holds for almost all other configurations, with only a few exceptions. For instance, considering a censoring rate of 70% and a true effect of $\beta = -0.25$ the bias of the NPPR model is for some configurations slightly larger than the one presented by the PPR model. If there is no treatment effect, that is, $\beta = 0$, the difference in biases mostly ranges in order of a magnitude of 0.004, which is very small. Only in case of a rather high amount of censoring, that is, 50%, and a small sample size of 50 participants this difference, given by 0.011, is noticeably larger. As expected, an application of Cox's PH model or the log-logistic PO model results in general in a larger bias, up to an absolute value of 0.495 (true value $\beta = 0.50$, 30% censoring, 500 participants) and 0.476 (true value $\beta = 0.50$, 30% censoring, 50 participants), respectively. In case of no censoring, the biases even exceed an absolute value of 0.500 multiple times for both models. If the true value of β is positive, both models tend to overestimate it, and underestimate it if the true value is negative. This is consistent with the trend observed for the HR and OR of the EU model over time (see Figure D in the supplementary material). Only in case of $\beta = 0$, the biases of all models are comparable in size, since with no effect present the PR, PH, and PO assumptions are fulfilled at the same time.

If the Weibull PH model is the true underlying model, the NPPR estimator has a slightly larger bias when trying to approximate the HR. It still never exceeds an absolute value of 0.138 if censoring is present, and an absolute value of 0.173 if there is no censoring. The negative values of β tend to be overestimated, while the positive ones are underestimated. Again, this behavior is consistent with the trend observed for the RR over time (see Figure 2). Similar results are observed if the log-logistic PO model is the true underlying model. The bias still never exceeds an absolute value of 0.223 and 0.259, respectively. Both Cox's PH and the log-logistic PO models performed adequate if the respective assumption is fulfilled. We note that the NPPR estimator is more precise for smaller censoring rates and a higher number of participants, as expected. Overall, it performs satisfactory and reliable in all cases under consideration.

Interestingly, all models tend to show a smaller bias for larger censoring rates if the respective assumption is violated. This most likely results from the censoring method, since later observations tend to be censored more often. Therefore, the events tend to concentrate on earlier time points for higher censoring rates. As seen in Figures 2 and D (see supplementary material), the RR, HR, and OR differ more at later time points. Consistent with this trend, the observed biases of all models are largest in case of no censoring if the respective assumption is violated.

3.3.2 | MSE

Tables 4–6 and B (see supplementary material) also depict the observed MSE. If the PPR model was assumed to be the true model, we note the same tendencies as seen with the bias. The NPPR estimator presents a smaller MSE for a low censoring rate of 30% in all cases. Considering 50% censoring it is overall smaller or equal to the one observed with the PPR model if $\beta \neq 0$ and for less than 500 participants if $\beta = 0.25$ or $\beta = -0.25$. The MSE of the NPPR estimator never exceeds 0.311 (true value $\beta = -0.50$, 70% censoring, 50 participants). On the other hand, the PPR model shows an MSE up to 0.600 (true value -0.50 , 30% censoring, 50 participants). Again, the observed MSEs of Cox's PH model and the log-logistic PO model are in general rather large for $\beta \neq 0$ since the PH/PO assumption is violated. However, even in case of $\beta = 0$, the NPPR outperforms these two models. If the Weibull PH model is the true underlying model, the NPPR estimator presents a similar MSE as for the PR case. It never exceeds 0.350 (true value $\beta = -0.50$, 70% censoring, 50 participants). Cox's PH model performs comparable indicating a slightly larger variance as the bias is generally smaller. If the log-logistic PO model is the true underlying model, the observed MSEs behave similarly. In this case, the MSE of the NPPR model never exceeds 0.363. If no censoring is present, the MSE of the NPPR model only exceeds 0.100 twice (PO data, 50 participants, true value $\beta = -0.5, 0.5$). Overall, we conclude that the MSEs corresponding to the NPPR model show promising behavior, also in case of a violation of the PR assumption.

3.3.3 | Coverage

As in general, the covariances $\text{cov}(\hat{\beta}_t, \hat{\beta}_{t'})$ for $t \neq t', t, t' \in \tilde{T}$, are unknown and it is therefore impossible to determine the variance of $\hat{\beta}$ in (2) directly, the CIs for the NPPR estimator were constructed according to the percentile bootstrap approach (Efron, 1981). Precisely, for each study with n participants, we drew an independent sample with replacement of size n from the simulated study. Consequently, the number of participants in each individual group is random and can differ from the one in the originally simulated study. The group assignment and status were not changed throughout. From this sample, $\hat{\beta}^*$ was estimated again using the NPPR estimator as described in Section 2. This procedure was repeated 500 times, yielding $\hat{\beta}^{*(1)}, \dots, \hat{\beta}^{*(500)}$. From these values, the empirical 2.5%-quantile and 97.5%-quantile, respectively, were determined, defining the corresponding 95% CI. For the PPR and the log-logistic PO models, we used the multivariate delta method (van der Vaart, 1998) to construct the CIs. CIs for Cox's PH model were directly extracted from output of the *coxph* function.

Details on the coverage for each simulated case are presented in the supplementary material (Tables C–F). In general, if the PPR model is the true underlying model, the simulated coverage of the CI for the NPPR estimator is very close to the desired confidence level of 95%. It rarely falls below this value, the smallest value is given by 93.4% and the highest coverage is given by 98.2%. The latter is reached in case of a true effect of -0.25 , 70% censoring, and 50 participants, which underlines the fact that CIs become rather conservative if the number of events is low.

Overall, the PPR model performs comparably. The smallest simulated coverage equals 89.4% and is observed in case of a true effect equal to 0.50 with a censoring rate of 30% and 100 participants. On the other hand, the smallest simulated coverage of Cox's PH model if censoring is present is equal to 80% ($\beta = -0.5$, 30% censoring, and 500 participants). The cases presenting a smaller coverage also present a larger bias. The coverage is lowest in case of no censoring (see Table F). Interestingly, it decreases with an increasing number of participants, most likely due to a more precise estimation of the HR, that is, close to the time-dependent RR but not equal to it. Therefore, the resulting bias and smaller variance combine to a lower coverage. However, in general, the coverage is very close to 95% especially for larger censoring rates. The coverage of the log-logistic PO model indicates a numerical problem since it often exceeds 99% and is never less than 94.4%. This is most likely caused by both the PO assumption and the parametric assumption being violated. If the Weibull PH model is the true underlying model, the coverage of the HR, approximated by the mean RR, is overall too low, which is a direct consequence of the violated PR assumption. However, we conclude that also in this case, for most of the

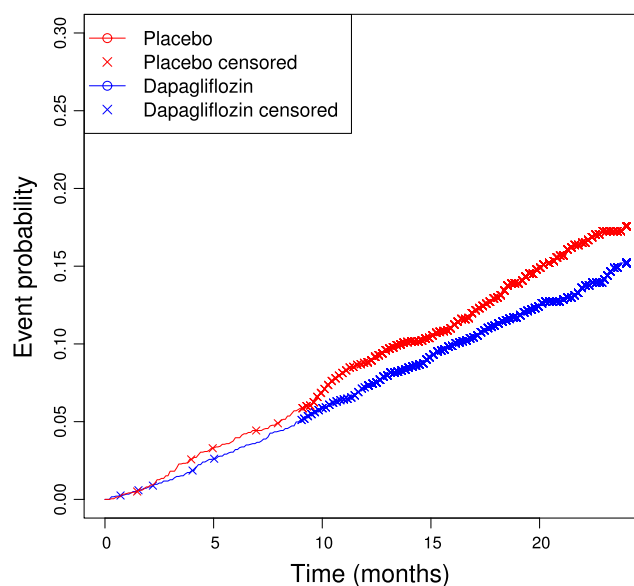


FIGURE 4 Event probability of the secondary outcome, all-cause mortality, and event time points during the DAPA-HF trial estimated with the Kaplan–Meier estimator.

configurations the approximation of the 95% level is still rather precise. If the log-logistic PO model is the true underlying model, the results are similar. The coverage of the OR is overall slightly smaller consistent with the bias being slightly larger. Both Cox’s PH and the log-logistic PO models perform as expected if the respective assumption is fulfilled.

3.3.4 | Numerical robustness

The NPPR estimator proves to be overall robust, independent of the true underlying model. Assuming the PPR model, failures (less than 12 in 1000 simulation runs) were only observed in five of the 45 scenarios. This is also true for the Weibull PH model and the log-logistic PO model as underlying model, where the NPPR model only failed a very few times (≤ 13), occurring only in case of a high censoring rate of 70% and a small sample size of 50 participants. Concerning the numerical robustness, the PPR model is clearly outperformed by the NPPR model, showing a higher robustness in almost every case. Cox’s PH and the log-logistic PO models both mostly present an even greater robustness. For the sake of brevity, details are deferred to the supplementary material (Tables G–J).

4 | CASE STUDY: DAPA-HF TRIAL

To illustrate the NPPR estimator, we use data from the DAPA-HF trial (McMurray et al., 2019). This randomized, double-blind, placebo-controlled trial evaluated dapagliflozin, a sodium glucose cotransporter-2 (SGLT-2) inhibitor, for reducing severe cardiovascular outcomes in patients with heart failure, and reduced ejection fraction. In 410 sites in 20 countries, 4744 patients were treated for a median observation time of 18.2 months. Here, we report the results for the trial’s secondary outcome, all-cause mortality, 274 (of 2373) patients experienced this outcome in the treatment group, but 324 (of 2371) in the control group. As we had no access to the original data, we digitized the Kaplan–Meier estimates from the original paper by the open software tool WebPlotDigitizer, version 3.8 (Rohatgi, 2015) and extracted the data by using the algorithms and R tools of Guyot et al. (2012). In order to validate this extraction process, we calculated an HR from the extracted data given by 0.837 (95% CI [0.712, 0.983]), which is essentially the same as the HR and its CI in the original paper, which was 0.83 (95% CI [0.71, 0.97]).

Figure 4 displays the estimated CDFs from the DAPA-HF trial. The assumption of PR seems to be reasonable. Using the method described in Section 2, we obtain $\hat{\beta}_{NPPR} = 0.178$ (95% CI [0.031, 0.346]), which corresponds to an RR of 0.837 (95% CI [0.708, 0.970]). The patients in the treatment group are therefore only 83.7% at risk of death for any cause relative to the patients in the placebo group. CIs were estimated using the bootstrap as described in Section 3.3.3. To visualize the

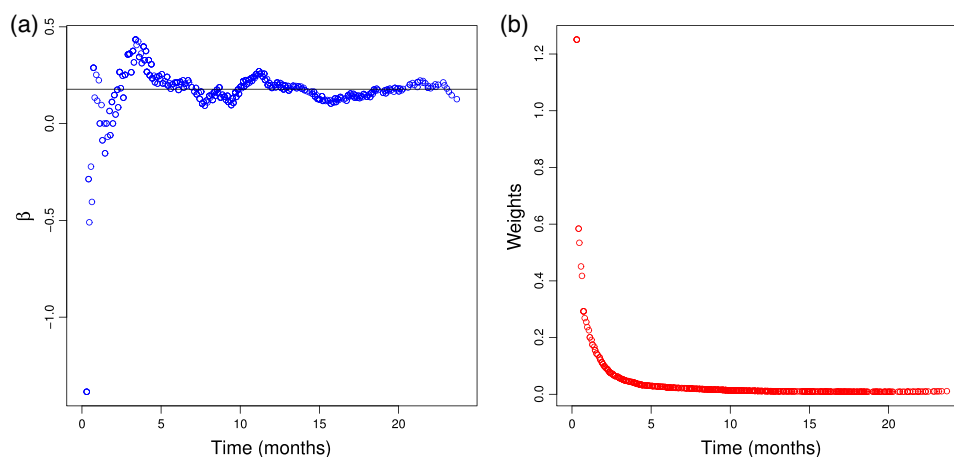


FIGURE 5 (a) Estimated $\hat{\beta}_t$ for the DAPA-HF trial at every event time point. The estimated $\hat{\beta}_{NPPR}$ overall is displayed by the horizontal line. (b) Weights at every event time point.

estimation process, Figure 5a shows the estimated $\hat{\beta}_t$ for each event time point. The weighted mean defining the NPPR estimator as defined in (2) is displayed by the solid line. The weights at each event time point are shown in Figure 5b.

The estimated NNT quickly decreases over time, likely due to the small number of reported events in the study. For instance, at the end of study ($t = 24$), only 29 (95% CI [19.256, 53.947]) patients treated with dapagliflozin would prevent one death relative to the placebo treatment. A detailed visualization of the NNT is shown in Figure O in the supplementary material.

5 | DISCUSSION AND CONCLUSION

In this paper, we proposed an NPPR model to assess a treatment effect in case of a two-group situation. Thereby, we solved the conceptional problems the HR and OR have shown in the past without losing the advantages of a nonparametric estimation method. By deducing the NNT from the model, we further provide an absolute measure in addition to the estimated RR. As no further specification beyond the PR assumption is necessary, the model is broadly applicable and provides a promising tool for the analysis of numerous applications, ranging from preclinical toxicology studies to late phase clinical trials, as, for instance RCTs. While not regarded in this paper, left truncated data could also be included (Tsai et al., 1987).

Using a bootstrap approach (Efron, 1981), we can construct CIs for the NPPR estimator and derive those of the estimated RR by multiplying with -1 and exponentiating. Due to the duality between CIs and statistical hypothesis tests, this further provides a test for the difference/ratio of the CDFs. To a certain degree, this holds true even if the PR assumption is violated since in this case the NPPR estimator evaluates the mean RR over time. Similar to the mean HR over time (Hernán, 2010) only regarding the mean RR from the beginning to the end of study might limit its interpretability. Therefore, again inspired by the mean HR (Hernán, 2010), we suggest to estimate the mean RR from the beginning up to multiple time points if the PR assumption is violated. However, if the estimated CDFs cross, the effect could be underestimated due to negative and positive $\hat{\beta}_t$ cancelling out and the method should be used with care. Hence, the NPPR estimator would highly benefit from the development of a goodness-of-fit test for the PR assumption.

Of course, the NPPR estimator comes along with some limitations. Up to this point, the model contains only one (binary) variable and it is not possible to include further variables in the model. A possible extension could be based on further stratification but might result in subgroup sample sizes that are too small for a proper evaluation of the Kaplan–Meier estimators. Moreover, the consideration of continuous variables is a challenging problem, which demands some future work. Another drawback, especially if compared to the parametric model, is the necessity of using bootstrap to calculate a CI as a formula for the variance of the estimator is not available yet. Alternatively, a different approach based on nonparametric maximum-likelihood estimation could circumvent these problems as well as supersede the calculation of weights.

Also, one important starting point of further research is the inclusion of competing risks, which is not possible as of now. A feasible way to address this issue without the need for a complicated construction might be to replace $1 - \hat{S}_i(t)$,

$i = 0, 1$, with the respective Aalen–Johansen estimator (Aalen & Johansen, 1978) since it is less susceptible to bias resulting from competing risks (Aalen & Johansen, 1978). However, in combination with an extension to multiple covariates, the incorporation of competing risks could offer an easier-to-interpret alternative (Andersen & Keiding, 2012) to the Fine–Gray subdistribution hazard model (Fine & Gray, 1999).

Another related topic would be a comparison of the PH and PR assumptions in terms of which is fulfilled more often in practice. This is a limitation of the given data example since the data do not seem to violate either.

We demonstrated that the NPPR estimator shows a very good performance if the PR assumption is fulfilled. For small to moderate censoring rates, it mostly outperforms the PPR model and is numerically more robust. The comparison with Cox's PH and the log-logistic models also shows that the direct estimation of the RR is advantageous over the approximation with the mean HR or the OR. If the PR assumption is not fulfilled it still shows satisfying behavior. However, if there are only few events, either due to high censoring rates or generally small sample sizes, problems can arise or even make an application impossible. In those cases, while still less robust, the PPR model can provide a better solution. The NPPR estimator, in general, estimates the mean RR over time. Therefore, applications outside of a PR scenario might also be of interest. Furthermore, similarly to the discussion about the PH assumption required from Cox's model, possible problems resulting from a violated PR assumption should be taken into consideration.

In summary, we strongly believe that the NPPR estimator is a useful addition to the existing tools for the analysis of time-to-event data, which not only circumvents the technical problems of the HR and the OR, respectively, but is also easy to interpret and comes along without an assumption on the underlying survival distribution.

ACKNOWLEDGMENTS

This work has been supported by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116).

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The R code, which can be used to reproduce the analysis of the case study and the simulation results, is available at <https://github.com/LuciaAmeis/NPPR-model-for-time-to-event-data>.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Lucia Ameis  <https://orcid.org/0009-0002-9420-6984>

Oliver Kuss  <https://orcid.org/0000-0003-3301-5869>

Annika Hoyer  <https://orcid.org/0000-0002-0241-9951>

Kathrin Möllenhoff  <https://orcid.org/0000-0001-7861-3892>

REFERENCES

- Aalen, O. O., & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3), 141–150. <http://www.jstor.org/stable/4615704>
- Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer. Corrected second printing, 1997. <https://doi.org/10.1007/978-1-4612-4348-9>
- Andersen, P. K., & Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11–12), 1074–1088. <https://doi.org/10.1002/sim.4385>
- Breslow, N., & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, 2(3), 437–453. <https://doi.org/10.1214/aos/1176342705>

- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374), 312–319. <https://doi.org/10.1080/01621459.1981.10477650>
- Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446), 496–509. <https://doi.org/10.1080/01621459.1999.10474144>
- Guyot, P., Ades, A. E., Ouwens, M. J. N. M., & Nicky, J. W. (2012). Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, 12, 9. <https://doi.org/10.1186/1471-2288-12-9>
- Hartung, J., Knapp, G., & Sinha, B. (2008). *Statistical meta-analysis with applications*. Wiley Series in Probability and Statistics. Wiley. <https://doi.org/10.1002/9780470386347>
- Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1), 13–15. Erratum in: *Epidemiology*, 22(1), 134, (2011). <https://doi.org/10.1097/EDE.0b013e3181c1ea43>
- Hildebrandt, M., Vervölgyi, E., & Bender, R. (2009). Calculation of NNTs in RCTs with time-to-event outcomes: A literature review. *BMC Medical Research Methodology*, 9(21). <https://doi.org/10.1186/1471-2288-9-21>
- Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8), 1–33. <https://doi.org/10.18637/jss.v070.i08>
- Jachno, K., Heritier, S., & Wolfe, R. (2019). Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. *BMC Medical Research Methodology*, 19(103). <https://doi.org/10.1186/s12874-019-0749-1>
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.2307/2281868>
- Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Wiley Series in Probability and Statistics. Wiley. <https://doi.org/10.1002/9781118032985>
- Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). Statistics for Biology and Health. Springer. <https://doi.org/10.1007/b97377>
- Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). Statistics for Biology and Health. Springer. <https://doi.org/10.1007/978-1-4419-6646-9>
- Kuss, O., & Hoyer, A. (2021). A proportional risk model for time-to-event analysis in randomized controlled trials. *Statistical Methods in Medical Research*, 30(2), 411–424. <https://doi.org/10.1177/0962280220953599>
- Lachin, J. M. (2010). *Biostatistical methods: The assessment of relative risks* (2nd ed.). Wiley Series in Probability and Statistics. Wiley. <https://doi.org/10.1002/9780470907412>
- Mandel, J. (1984). *The statistical analysis of experimental data*. Dover Publications.
- McMurray, J. J. V., Solomon, S. D., Inzucchi, S. E., Køber, L., Kosiborod, M. N., Martinez, F. A., Ponikowski, P., Sabatine, M. S., Anand, I. S., Bělohávek, J., Böhm, M., Chiang, C.-E., Chopra, V. K., de Boer, R. A., Desai, A. S., Diez, M., Drozd, J., Dukát, A., Ge, J., ... Langkilde, A.-M. (2019). Dapagliflozin in patients with heart failure and reduced ejection fraction. *New England Journal of Medicine*, 381(21), 1995–2008. <https://doi.org/10.1056/NEJMoa1911303>
- NCBI. (1996). PubMed. <https://pubmed.ncbi.nlm.nih.gov/?term=time-to-event>
- Oehlert, G. W. (1992). A note on the delta method. *American Statistician*, 46(1), 27–29. <https://doi.org/10.2307/2684406>
- Porta, M. (2016). *A dictionary of epidemiology*. Oxford University Press. Current online version. <https://doi.org/10.1093/acref/9780199976720.001.0001>
- Rohatgi, A. (2015). WebPlotDigitizer. <https://automeris.io/WebPlotDigitizer/>
- Rosner, B. (2016). *Fundamentals of biostatistics* (8th ed.). Cengage Learning.
- Royston, P., & Parmar, M. K. (2014). An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials*, 15(314). <https://doi.org/10.1186/1745-6215-15-314>
- Schwartz, L. M., Woloshin, S., & Welch, H. G. (1999). Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *New England Journal of Medicine*, 341(4), 279–283. <https://doi.org/10.1056/NEJM199907223410411>
- Sistrom, C. L., & Garvan, C. W. (2004). Proportions, odds, and risk. *Radiology*, 230(1), 12–19. <https://doi.org/10.1148/radiol.2301031028>
- Struthers, C. A., & Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2), 363–369. <https://doi.org/10.2307/2336212>
- Sutradhar, R., & Austin, P. C. (2018). Relative rates not relative risks: Addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology*, 28(1), 54–57. <https://doi.org/10.1016/j.annepidem.2017.10.014>
- The Academy of Medical Sciences. (2017). *Sources of evidence for assessing the safety, efficacy and effectiveness of medicines*. <https://acmedsci.ac.uk/file-download/86466482>
- Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer. <https://doi.org/10.1007/978-1-4757-3294-8>
- Therneau, T. M. (2023). *A package for survival analysis in R*. R Package version 3.5-7. <https://CRAN.R-project.org/package=survival>
- Tsai, W., Nicholas, P. J., & Wang, M. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4), 883–886. <https://doi.org/10.2307/2336484>
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802256>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ameis, L., Kuss, O., Hoyer, A., & Möllenhoff, K. (2024). A nonparametric proportional risk model to assess a treatment effect in time-to-event data. *Biometrical Journal*, 66, 2300147.

<https://doi.org/10.1002/bimj.202300147>

Supplementary material to: A non-parametric proportional risk model to assess a treatment effect in time-to-event data

L. Ameis^{1,2}, O. Kuss³, A. Hoyer⁴, and K. Möllenhoff^{1,2}

¹Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne, Cologne, Germany

²Mathematical Institute, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

³German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometrics and Epidemiology, Düsseldorf, Germany

⁴Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Bielefeld, Germany

In the following supplementary material additional tables and further visualizations of the conducted analyses are provided. The order of presentation is in line with the main text.

2 Methodology

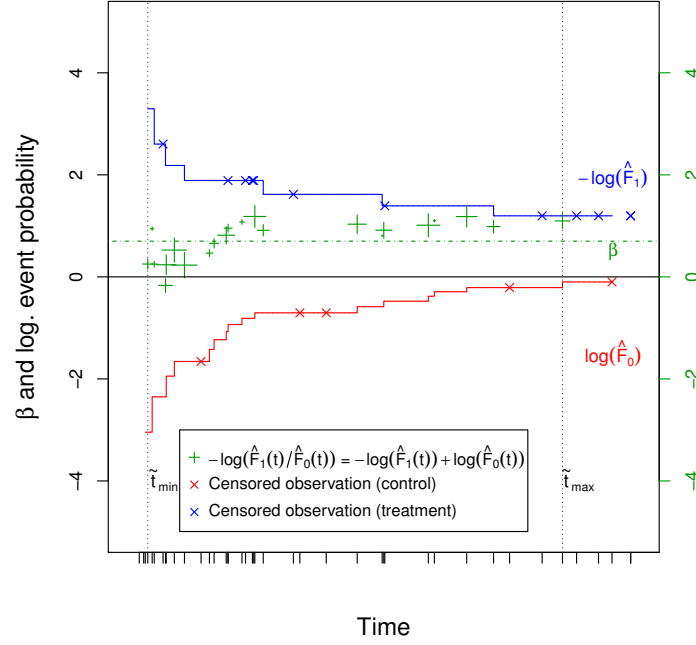


Figure A: Visualization of the estimation method on the logarithmic scale. The size of the pluses symbolises the weight given to respective estimated RR at this time point. The larger the plus, the larger the weight.

3 Simulation study

3.0 Calculations

Lemma 3.1. *The RR of the PPR model is defined by*

$$RR_{PPR} = \left(\frac{\theta_1}{\theta_0} \right)^\alpha.$$

The HR and OR are both defined by

$$HR_{PPR} = OR_{PPR} = k^\alpha \left(\frac{1 - (\theta_0 t)^\alpha}{1 - (k\theta_0 t)^\alpha} \right), \text{ for } t \leq \min \left\{ \frac{1}{\theta_1}, \frac{1}{\theta_0} \right\}.$$

Note, that the RR does not depend on t but the HR and the OR do.

Proof. First, we will calculate the RR as follows:

$$RR_{PPR} = \frac{F_{EU,1}(t)}{F_{EU,0}(t)} = \frac{(\theta_1 t)^\alpha}{(\theta_0 t)^\alpha} = \frac{\theta_1^\alpha t^\alpha}{\theta_0^\alpha t^\alpha} = \left(\frac{\theta_1}{\theta_0} \right)^\alpha$$

To calculate the HR and the OR we will next calculate the formulas for the distribution function and hazard function. The distribution function is derived from the CDF and has the form

$$f_{EU,i} = \alpha \theta_i^\alpha t^{\alpha-1}, \text{ for } t < \frac{1}{\theta_i}, \quad i = 0, 1.$$

Therefore, the hazard function is defined by

$$h_{EU,i}(t) = \frac{f_{EU,i}(t)}{1 - F_{EU,i}(t)} = \frac{\alpha \theta_i^\alpha t^{\alpha-1}}{1 - (\theta_i t)^\alpha}.$$

With this we can now determine the HR:

$$HR_{PPR}(t) = \frac{h_{EU,1}(t)}{h_{EU,0}(t)} = \frac{\frac{\alpha(\theta_1)^\alpha t^{\alpha-1}}{1 - (\theta_1 t)^\alpha}}{\frac{\alpha\theta_0^\alpha t^{\alpha-1}}{1 - (\theta_0 t)^\alpha}} = \left(\frac{\theta_1}{\theta_0}\right)^\alpha \left(\frac{1 - (\theta_0 t)^\alpha}{1 - (\theta_1 t)^\alpha}\right)$$

The OR is defined by:

$$OR_{PPR}(t) = \frac{\frac{F_1(t)}{1 - F_1(t)}}{\frac{F_0(t)}{1 - F_0(t)}} = \frac{\frac{(\theta_1 t)^\alpha}{1 - (\theta_1 t)^\alpha}}{\frac{(\theta_0 t)^\alpha}{1 - (\theta_0 t)^\alpha}} = \left(\frac{\theta_1}{\theta_0}\right)^\alpha \left(\frac{1 - (\theta_0 t)^\alpha}{1 - (\theta_1 t)^\alpha}\right)$$

□

Remark 3.2. The survival function of Cox's PH model in the RCT case is defined by $S_0(t)^{\exp(\beta X_i)}$ for $t > 0$ with an unspecified baseline survival function $S_0(t)$ and $X_i = \begin{cases} 1 & \text{treatment} \\ 0 & \text{control} \end{cases}$.

Therefore, the CDF is given by $F_{Cox,i}(t) = 1 - S_0(t)^{\exp(\beta X_i)}$ and the RR by

$$\frac{F_{1,Cox}(t)}{F_{0,Cox}(t)} = \frac{1 - S_0(t)^{\exp(\beta)}}{1 - S_0(t)^{\exp(0)}} = \frac{1 - S_0(t)^{\exp(\beta)}}{1 - S_0(t)}.$$

In general, this formula depends on t . For example, setting $S_0(t) = 1 - (\theta t)^\alpha$ as exponentiated-uniform distributed the RR is

$$\frac{F_{1,Cox-EU}(t)}{F_{0,Cox-EU}(t)} = \frac{1 - (1 - (\theta t)^\alpha)^{\exp(\beta)}}{(\theta t)^\alpha}.$$

Inspired by the simulation study (Section 3, Table 3) we set $\alpha = 0.859$, $\theta = 0.009$ and $\beta = 0.5$. Figure B shows the RR over time.

Lemma 3.3. The OR of the log-logistic PO model is defined by

$$OR_{LL} = \left(\frac{b_0}{b_1}\right)^a.$$

¹Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health Springer New York, NY. 3rd ed. 2012. doi: 10.1007/978-1-4419-6646-9

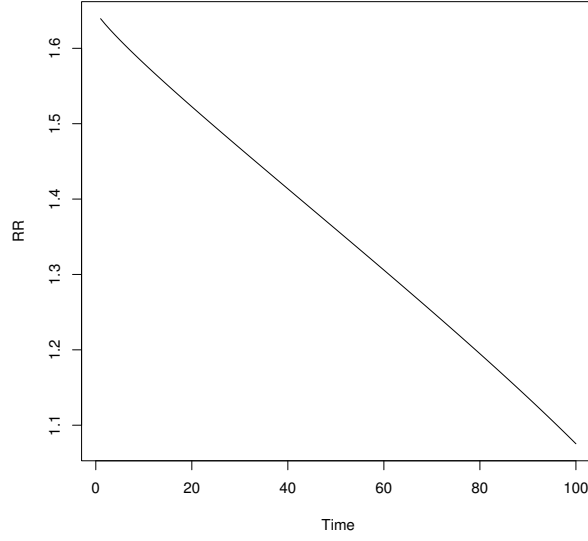


Figure B: RR of Cox's PH model with an exponentiated-uniform distributed baseline survival function and $\alpha = 0.895$, $\theta = 0.009$ and $\beta = 0.5$ over time.

The RR and the HR are both defined by

$$RR_{LL} = HR_{LL} = \frac{b_0^a + t^a}{b_1^a + t^a} \text{ for } t > 0.$$

Note, that the RR and the HR both depend on t .

Proof. The CDF of log-logistic PO model is defined by

$$F_{LL,i}(t) = 1 - \frac{1}{1 + \left(\frac{t}{b_i}\right)^a} \text{ for } i = 0, 1$$

and therefore, we can calculate the OR:

$$\begin{aligned}
OR_{LL} &= \frac{\frac{F_{LL,1}(t)}{1-F_{LL,1}(t)}}{\frac{F_{LL,0}(t)}{1-F_{LL,0}(t)}} = \frac{\frac{1 - \frac{1}{1 + \left(\frac{t}{b_1}\right)^a}}{\frac{1}{1 + \left(\frac{t}{b_1}\right)^a}}}{\frac{1 - \frac{1}{1 + \left(\frac{t}{b_0}\right)^a}}{\frac{1}{1 + \left(\frac{t}{b_0}\right)^a}}} \\
&= \frac{1 - \frac{1}{1 + \left(\frac{t}{b_1}\right)^a}}{\frac{1}{1 + \left(\frac{t}{b_1}\right)^a}} \frac{\frac{1}{1 + \left(\frac{t}{b_0}\right)^a}}{1 - \frac{1}{1 + \left(\frac{t}{b_0}\right)^a}} \\
&= \frac{\frac{\left(\frac{t}{b_1}\right)^a}{1 + \left(\frac{t}{b_1}\right)^a}}{\frac{1}{1 + \left(\frac{t}{b_1}\right)^a}} \frac{\frac{1}{1 + \left(\frac{t}{b_0}\right)^a}}{\frac{\left(\frac{t}{b_0}\right)^a}{1 + \left(\frac{t}{b_0}\right)^a}} \\
&= \frac{\frac{\left(\frac{t}{b_1}\right)^a}{\left(1 + \left(\frac{t}{b_1}\right)^a\right)\left(1 + \left(\frac{t}{b_0}\right)^a\right)}}{\frac{\left(\frac{t}{b_0}\right)^a}{\left(1 + \left(\frac{t}{b_1}\right)^a\right)\left(1 + \left(\frac{t}{b_0}\right)^a\right)}} \\
&= \frac{\left(\frac{t}{b_1}\right)^a}{\left(\frac{t}{b_0}\right)^a} \\
&= \left(\frac{b_0}{b_1}\right)^a.
\end{aligned}$$

The RR is defined by

$$\begin{aligned}
RR_{LL} &= \frac{F_{LL,1}(t)}{F_{LL,0}(t)} = \frac{1 - \frac{1}{1 + \left(\frac{t}{b_1}\right)^a}}{1 - \frac{1}{1 + \left(\frac{t}{b_0}\right)^a}} \\
&= \frac{\frac{\left(\frac{t}{b_1}\right)^a}{1 + \left(\frac{t}{b_1}\right)^a}}{\frac{\left(\frac{t}{b_0}\right)^a}{1 + \left(\frac{t}{b_0}\right)^a}} \\
&= \frac{\left(\frac{t}{b_1}\right)^a}{1 + \left(\frac{t}{b_1}\right)^a} \frac{1 + \left(\frac{t}{b_0}\right)^a}{\left(\frac{t}{b_0}\right)^a} \\
&= \frac{\left(\frac{t}{b_1}\right)^a + \left(\frac{t^{2a}}{b_0^a b_1^a}\right)}{\left(\frac{t}{b_0}\right)^a + \left(\frac{t^{2a}}{b_0^a b_1^a}\right)} \\
&= \frac{\left(\frac{b_0 t}{b_0^a b_1^a} + t^{2a}\right)}{\left(\frac{b_1 t}{b_0^a b_1^a} + t^{2a}\right)} \\
&= \left(\frac{(b_0 t)^a + t^{2a}}{b_0^a b_1^a}\right) \left(\frac{b_0^a b_1^a}{(b_1 t)^a + t^{2a}}\right) \\
&= \frac{b_0^a + t^a}{b_1^a + t^a}
\end{aligned}$$

To calculate the HR we first derive the distribution function and the hazard function. It is

$$\begin{aligned}
f_{LL,i}(t) &= \frac{-d \left(1 + \left(\frac{t}{b_i}\right)^a\right)^{-1}}{dt} \\
&= \frac{\frac{at^{a-1}}{b_i^a}}{\left(1 + \left(\frac{t}{b_i}\right)^a\right)^2}
\end{aligned}$$

and therefore

$$h_{LL,i}(t) = \frac{f_{LL,i}(t)}{1 - F_{LL,i}(t)} = \frac{\frac{\frac{at^{a-1}}{b_i^a}}{\left(1 + \left(\frac{t}{b_i}\right)^a\right)^2}}{\frac{1}{1 + \left(\frac{t}{b_i}\right)^a}} = \frac{\frac{at^{a-1}}{b_i^a}}{\left(1 + \left(\frac{t}{b_i}\right)^a\right)}.$$

Now, it is

$$\begin{aligned}
HR_{LL} &= \frac{h_{LL,1}(t)}{h_{LL,0}(t)} = \frac{\frac{at^{a-1}}{b_1^a}}{\frac{at^{a-1}}{b_0^a}} \frac{\left(1 + \left(\frac{t}{b_1}\right)^a\right)}{\left(1 + \left(\frac{t}{b_0}\right)^a\right)} \\
&= \frac{\frac{at^{a-1}}{b_1^a}}{\frac{at^{a-1}}{b_0^a}} \frac{\left(1 + \left(\frac{t}{b_0}\right)^a\right)}{\left(1 + \left(\frac{t}{b_1}\right)^a\right)} \\
&= \frac{at^{a-1}}{b_1^a} \frac{b_0^a}{at^{a-1}} \frac{\left(1 + \left(\frac{t}{b_0}\right)^a\right)}{\left(1 + \left(\frac{t}{b_1}\right)^a\right)} \\
&= \frac{b_0^a}{b_1^a} \frac{\left(1 + \left(\frac{t}{b_0}\right)^a\right)}{\left(1 + \left(\frac{t}{b_1}\right)^a\right)} \\
&= \frac{b_0^a + t^a}{b_1^a + t^a}
\end{aligned}$$

□

Lemma 3.4. *The HR of the Weibull PH model is defined by*

$$HR_{Weibul} = \left(\frac{\lambda_0}{\lambda_1}\right)^k.$$

The RR is defined by

$$RR_{Weib} = \frac{1 - e^{-\left(\frac{t}{\lambda_1}\right)^k}}{1 - e^{-\left(\frac{t}{\lambda_0}\right)^k}} \text{ for } t > 0.$$

and the OR by

$$OR_{Weib} = \frac{1 - e^{-\left(\frac{t}{\lambda_1}\right)^k}}{1 - e^{-\left(\frac{t}{\lambda_0}\right)^k}} e^{-t^k \left(\frac{\lambda_1 + \lambda_0}{\lambda_0 \lambda_1}\right)^k} \text{ for } t > 0.$$

Note, that the RR and the OR both depend on t.

Proof. The CDF of the Weibull PH model is defined by

$$F_{Weib,i}(t) = 1 - e^{-\left(\frac{t}{\lambda_i}\right)^k} \text{ for } i = 0, 1$$

and therefore we can derive the distribution function:

$$f_{Weib,i}(t) = \frac{k}{\lambda_i} \left(\frac{t}{\lambda_i}\right)^{k-1} e^{-\left(\frac{t}{\lambda_i}\right)^k}$$

Therefore, the hazard function is

$$h_{Weib,i}(t) = \frac{f_{Weib,i}(t)}{1 - F_{Weib,i}(t)} = \frac{\frac{k}{\lambda_i} \left(\frac{t}{\lambda_i}\right)^{k-1} e^{-\left(\frac{t}{\lambda_i}\right)^k}}{e^{-\left(\frac{t}{\lambda_i}\right)^k}} = \frac{k}{\lambda_i} \left(\frac{t}{\lambda_i}\right)^{k-1}.$$

Now the HR can be determined as follows:

$$HR_{Weib} = \frac{h_{Weib,1}(t)}{h_{Weib,0}(t)} = \frac{\frac{k}{\lambda_1} \left(\frac{t}{\lambda_1}\right)^{k-1}}{\frac{k}{\lambda_0} \left(\frac{t}{\lambda_0}\right)^{k-1}} = \left(\frac{\lambda_0}{\lambda_1}\right)^k$$

The RR is defined by

$$RR_{Weib} = \frac{F_{Weib,1}(t)}{F_{Weib,0}(t)} = \frac{1 - e^{-\left(\frac{t}{\lambda_1}\right)^k}}{1 - e^{-\left(\frac{t}{\lambda_0}\right)^k}}$$

and the OR by

$$\begin{aligned} OR_{Weib} &= \frac{\frac{F_{LL,1}(t)}{1-F_{LL,1}(t)}}{\frac{F_{LL,0}(t)}{1-F_{LL,0}(t)}} = \frac{\frac{1-e^{-\left(\frac{t}{\lambda_1}\right)^k}}{e^{-\left(\frac{t}{\lambda_1}\right)^k}}}{\frac{1-e^{-\left(\frac{t}{\lambda_0}\right)^k}}{e^{-\left(\frac{t}{\lambda_0}\right)^k}}} \\ &= \frac{1 - e^{-\left(\frac{t}{\lambda_1}\right)^k}}{1 - e^{-\left(\frac{t}{\lambda_0}\right)^k}} \frac{e^{-\left(\frac{t}{\lambda_0}\right)^k}}{e^{-\left(\frac{t}{\lambda_1}\right)^k}} \\ &= \frac{1 - e^{-\left(\frac{t}{\lambda_1}\right)^k}}{1 - e^{-\left(\frac{t}{\lambda_0}\right)^k}} e^{-\left(\frac{t}{\lambda_0}\right)^k} e^{\left(\frac{t}{\lambda_1}\right)^k} \\ &= \frac{1 - e^{-\left(\frac{t}{\lambda_1}\right)^k}}{1 - e^{-\left(\frac{t}{\lambda_0}\right)^k}} e^{-t^k \left(\frac{\lambda_1 + \lambda_0}{\lambda_0 \lambda_1}\right)^k}. \end{aligned}$$

□

3.1 Simulation

3.1.1 Analysis of the DAPA-HF trial with focus on its primary outcome

Here we report the results for the DAPA-HF trial's primary outcome, time to worsening heart failure or cardiovascular death. Overall, 385 (out of 2,373) patients experienced this outcome in the treatment group, but 500 (out of 2371) in the control group. The data was extracted as described in Section 4. In order to validate this extraction process, we calculated a hazard ratio (with 95% CI) from the extracted data given by 0.745 [0.653, 0.851], which is essentially the same as the hazard ratio and its CI in the original paper, which was 0.74 [0.65, 0.85]. Figure a) depicts the Kaplan-Meier estimates from the DAPA-HF trail. The PR assumption seems to be reasonable. Using the NPPR estimator as described in Section 2 we obtain $\hat{\beta}_{NPPR} = 0.320$ (95%CI [0.178, 0.454]), which corresponds to a RR of 0.726 (95%CI [0.635, 0.837]). The patients in the treatment group are therefore only 72.6% at risk of worsening heart failure or cardiovascular death relative to the patients in the placebo group. Again, confidence intervals were estimated using the bootstrap as described in Section 3.3.3. Figure b) shows the estimated $\hat{\beta}_t$ for each time point. The estimated $\hat{\beta}_{NPPR}$ is displayed by the black line. Figure c) depicts the weights at every time point used to calculate the weighted mean of the $\hat{\beta}_t$ as described in (2). Figure d) depicts the number needed to treat over time. It quickly decreases, likely due to the small number of reported events in the study. At the end of study ($t = 24$) it requires only 15 (14.463 95%CI [10.214, 24.313]) patients treated with dapagliflozin would prevent one death relative to the placebo treatment.

Note that since there is also death from other causes than worsening heart failure or cardiovascular death, this analysis might be biased. However, this does not influence the results of the simulation study which is based on assuming only death of all causes as an event.

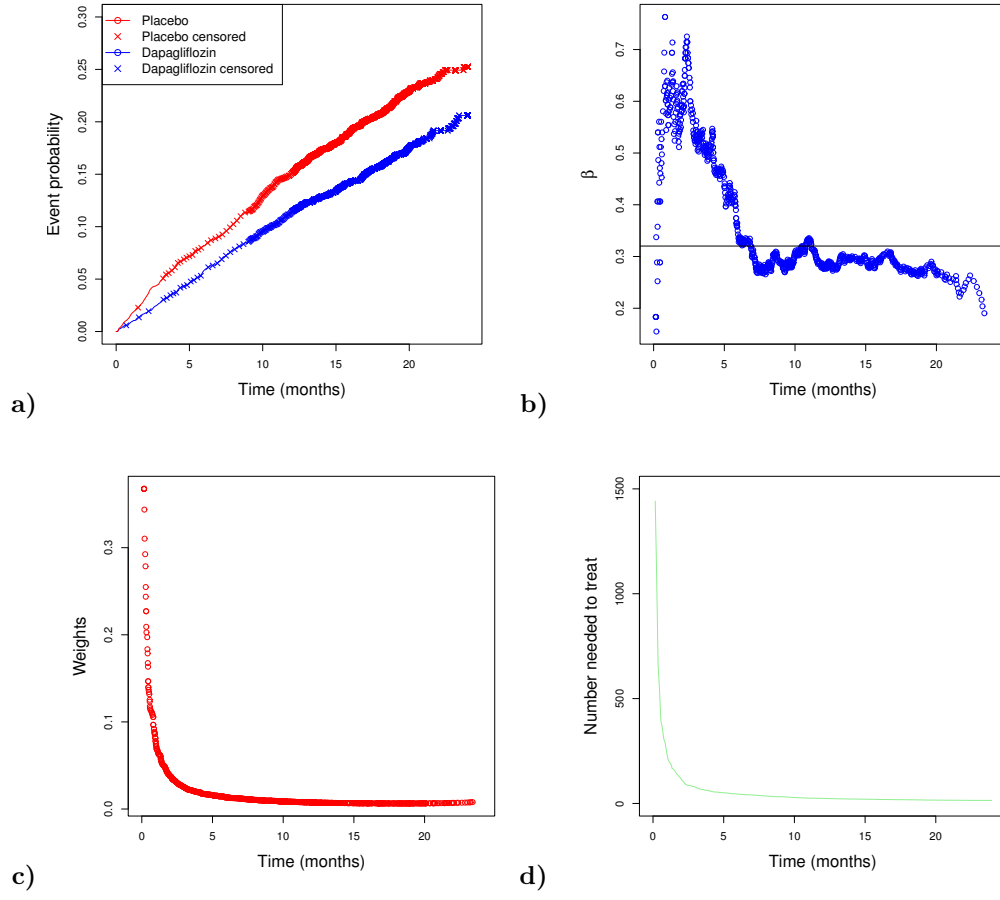


Figure C: Figure a) depicts the Kaplan-Meier estimates from the DAPA-HF trial with the primary outcome. Figure b) shows the estimated $\hat{\beta}_t$. Figure c) depicts the weights of the $\hat{\beta}_t$. Figure d) depicts the NNT over time.

3.1.2 Simulation Scenarios

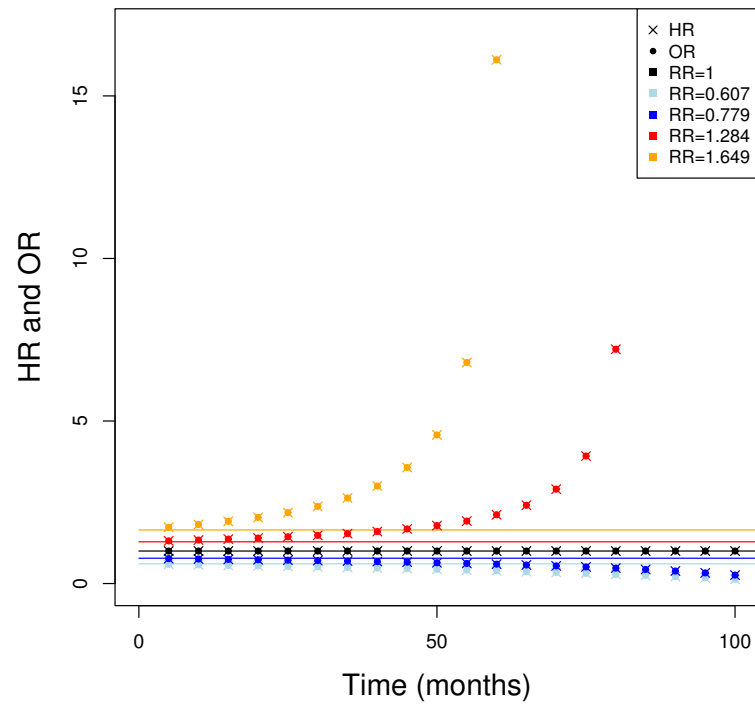


Figure D: The HR and OR of the PPR model over time in comparison with the constant value indicated by the straight lines.

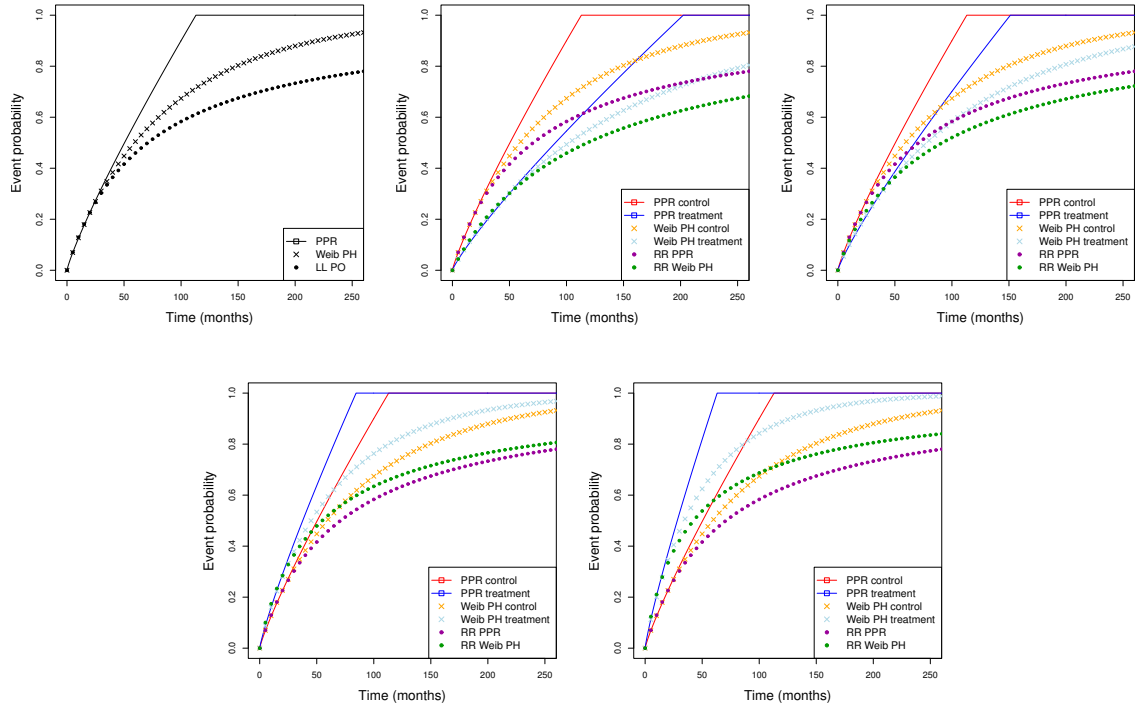


Figure E: CDFs of the PPR model in comparison with the CDFs of the Weibull PH model and the log-logistic PO model for true underlying RR/HR/OR of 1 (top left), 0.607 (top center), 0.779 (top right), 1.284 (lower left) and 1.649 (lower right).

Table A: Parameters of the uniform distribution corresponding to the proportion of censoring for all different cases.

a) PPR

Number of censored participants (%)	True underlying effect β	Parameters UD	
		min	max
30	0.00	0	175
30	0.50	0	250
30	0.25	0	200
30	-0.25	0	150
30	-0.50	0	130
50	0.00	0	100
50	0.50	0	140
50	0.25	0	125
50	-0.25	0	90
50	-0.50	0	80
70	0.00	0	60
70	0.50	0	75
70	0.25	0	70
70	-0.25	0	50
70	-0.50	0	40

b) Weibull PH

Number of censored participants (%)	True underlying effect β	Parameters UD	
		min	max
30	0.00	0	275
30	0.50	0	350
30	0.25	0	320
30	-0.25	0	250
30	-0.50	0	220
50	0.00	0	130
50	0.50	0	150
50	0.25	0	155
50	-0.25	0	115
50	-0.50	0	110
70	0.00	0	60
70	0.50	0	75
70	0.25	0	70
70	-0.25	0	50
70	-0.50	0	45

c) Log-Logistic PO

Number of censored participants (%)	True underlying effect β	Parameters UD	
		min	max
30	0.00	0	500
30	0.50	0	650
30	0.25	0	550
30	-0.25	0	400
30	-0.50	0	400
50	0.00	0	175
50	0.50	0	225
50	0.25	0	200
50	-0.25	0	150
50	-0.50	0	125
70	0.00	0	60
70	0.50	0	75
70	0.25	0	75
70	-0.25	0	55
70	-0.50	0	50

3.3 Results

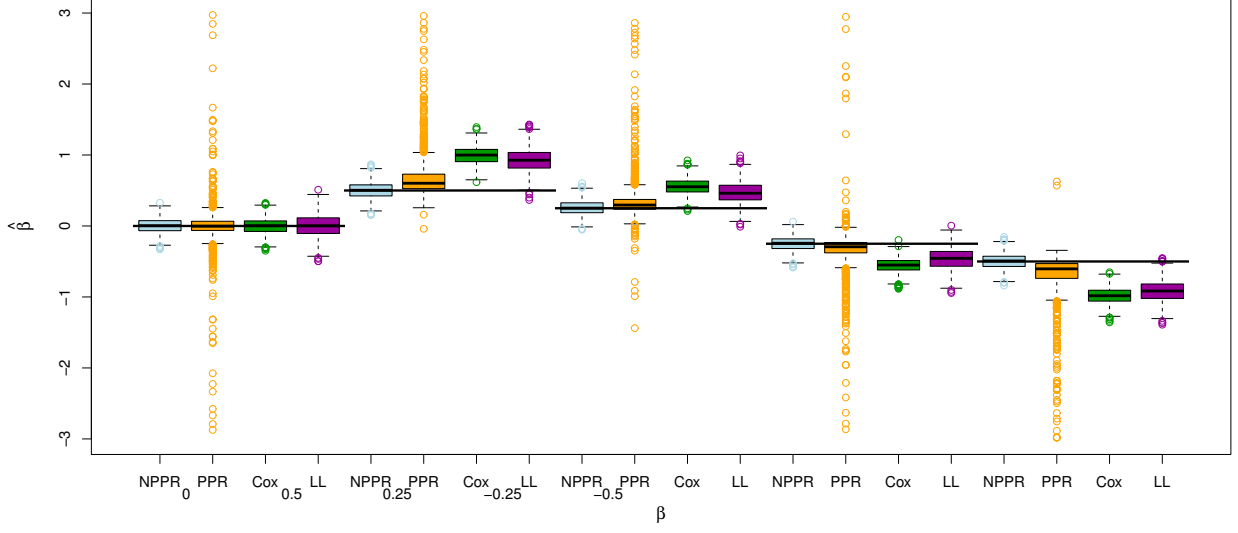


Figure F: Comparison of boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR), $-\log(\widehat{RR}_{PPR})$ (PPR), $-\log(\widehat{HR}_{Cox})$ (Cox) and $-\log(\widehat{OR}_{LL})$ (LL) for all different true β if the PPR model is the underlying model, each in case of 500 participants and 30% censoring rate. Horizontal bars indicate the true underlying effect.

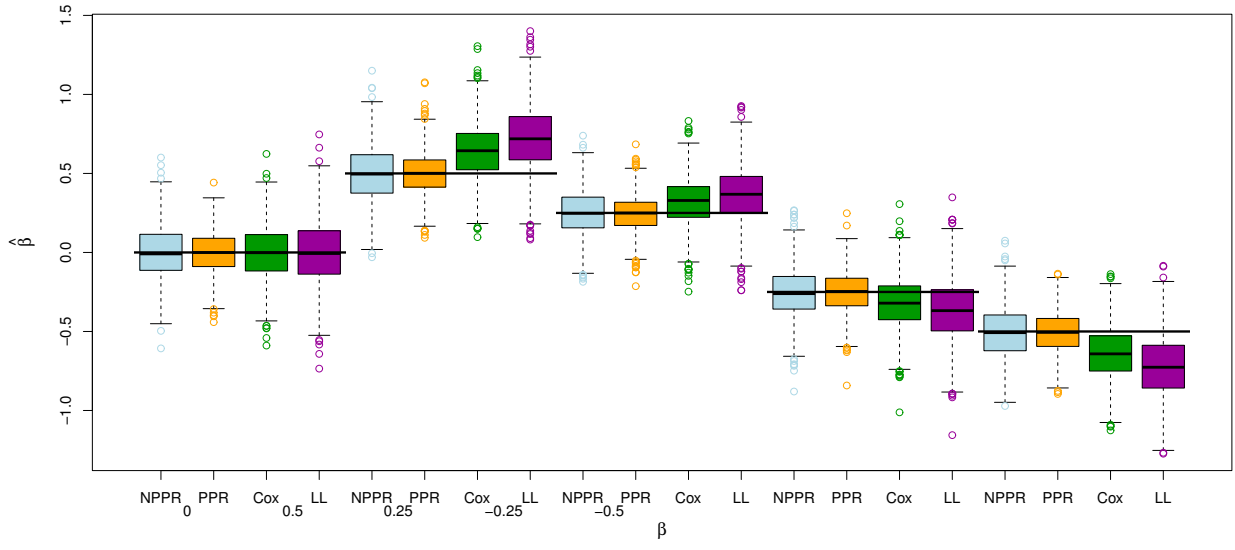


Figure G: Comparison of boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR), $-\log(\widehat{RR}_{PPR})$ (PPR), $-\log(\widehat{HR}_{Cox})$ (Cox) and $-\log(\widehat{OR}_{LL})$ (LL) for all different true β if the PPR model is the underlying model, each in case of 500 participants and 70% censoring rate. Horizontal bars indicate the true underlying effect.

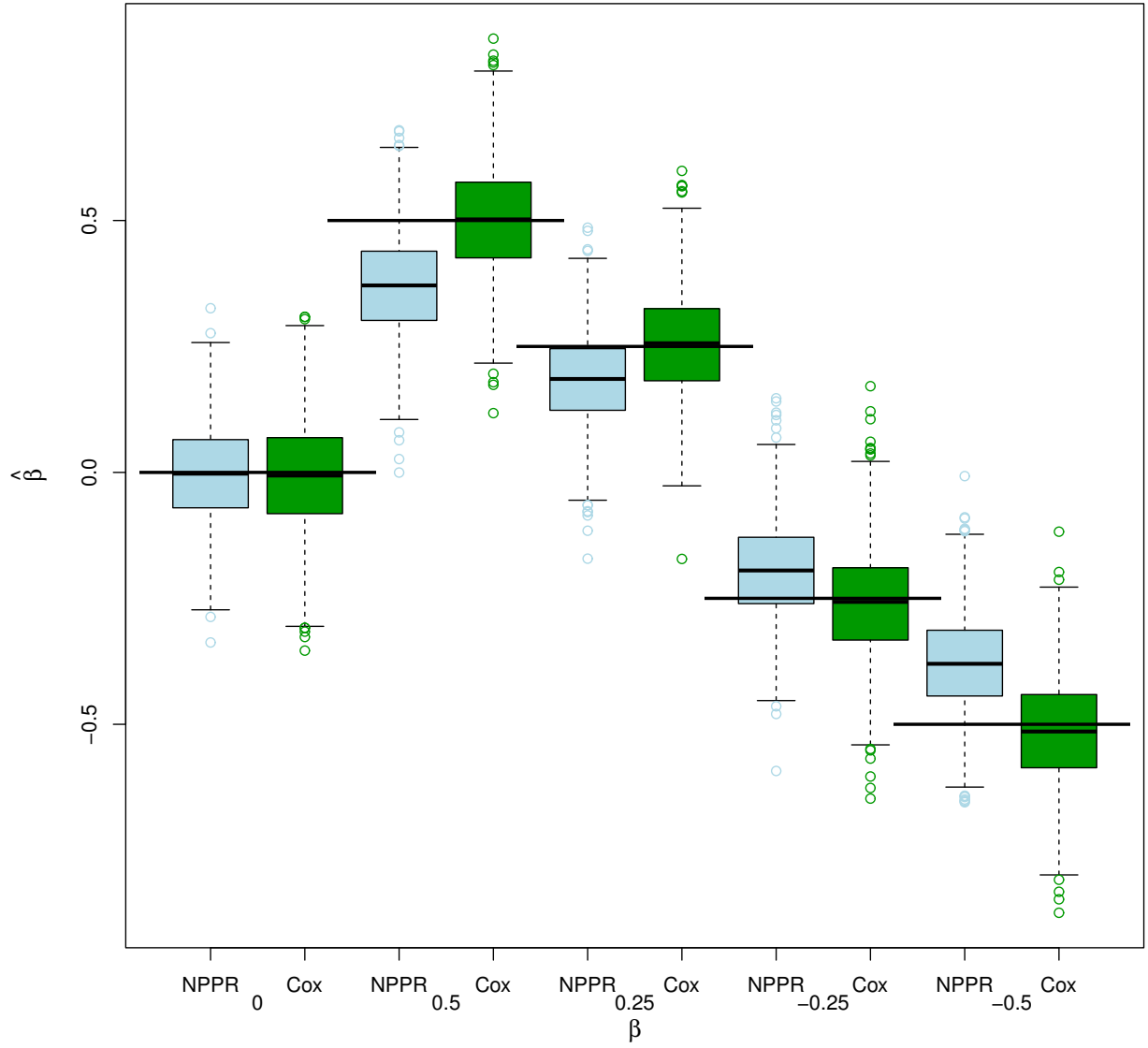


Figure H: Boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR) and $-\log(\widehat{HR}_{Cox})$ (Cox) for all different true β if the Weibull PH model is the underlying model, each in case of 500 participants and 30% censoring rate. Horizontal bars indicate the true underlying effect.

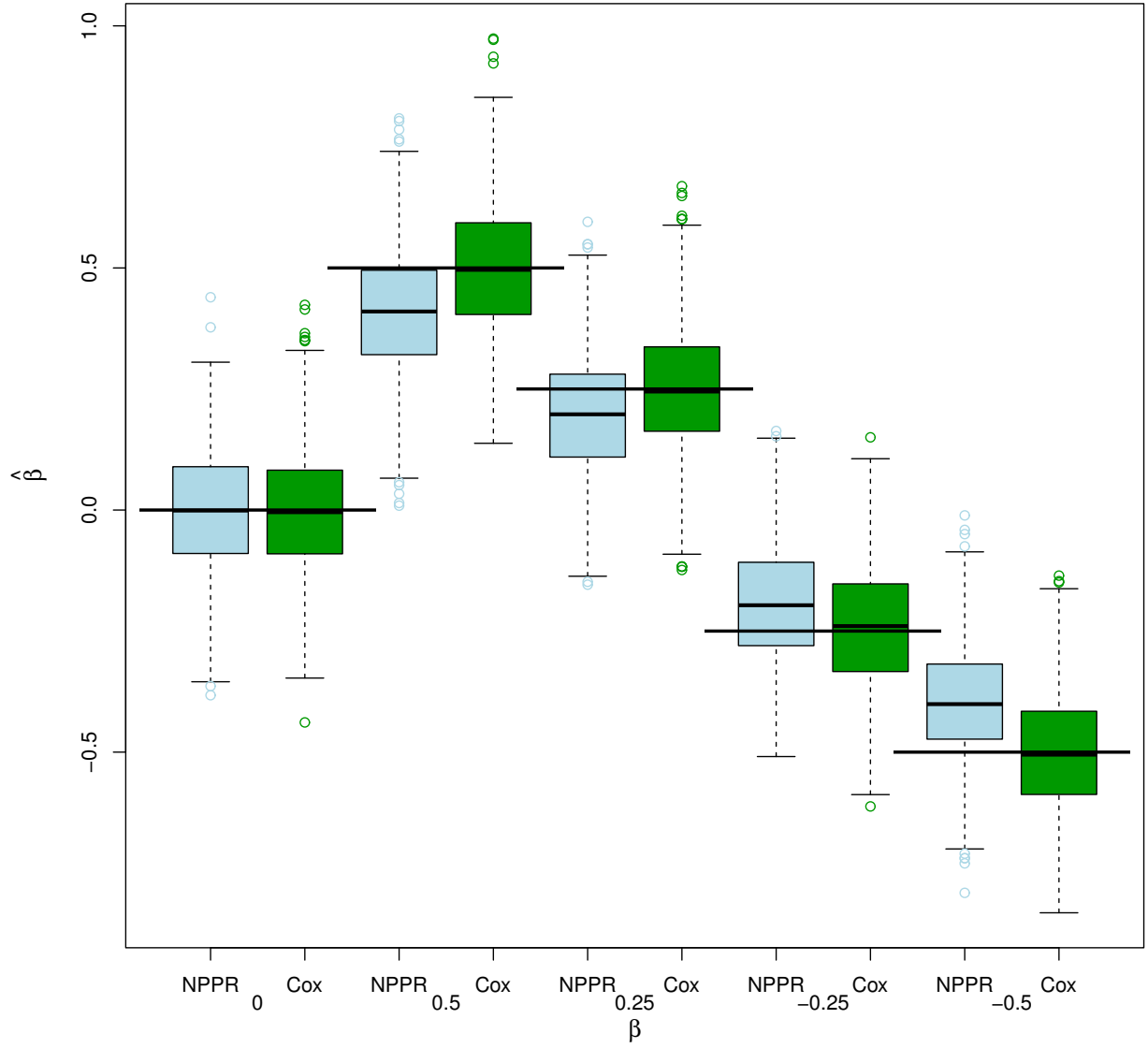


Figure I: Boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR) and $-\log(\widehat{HR}_{Cox})$ (Cox) for all different true β if the Weibull PH model is the underlying model, each in case of 500 participants and 50% censoring rate. Horizontal bars indicate the true underlying effect.

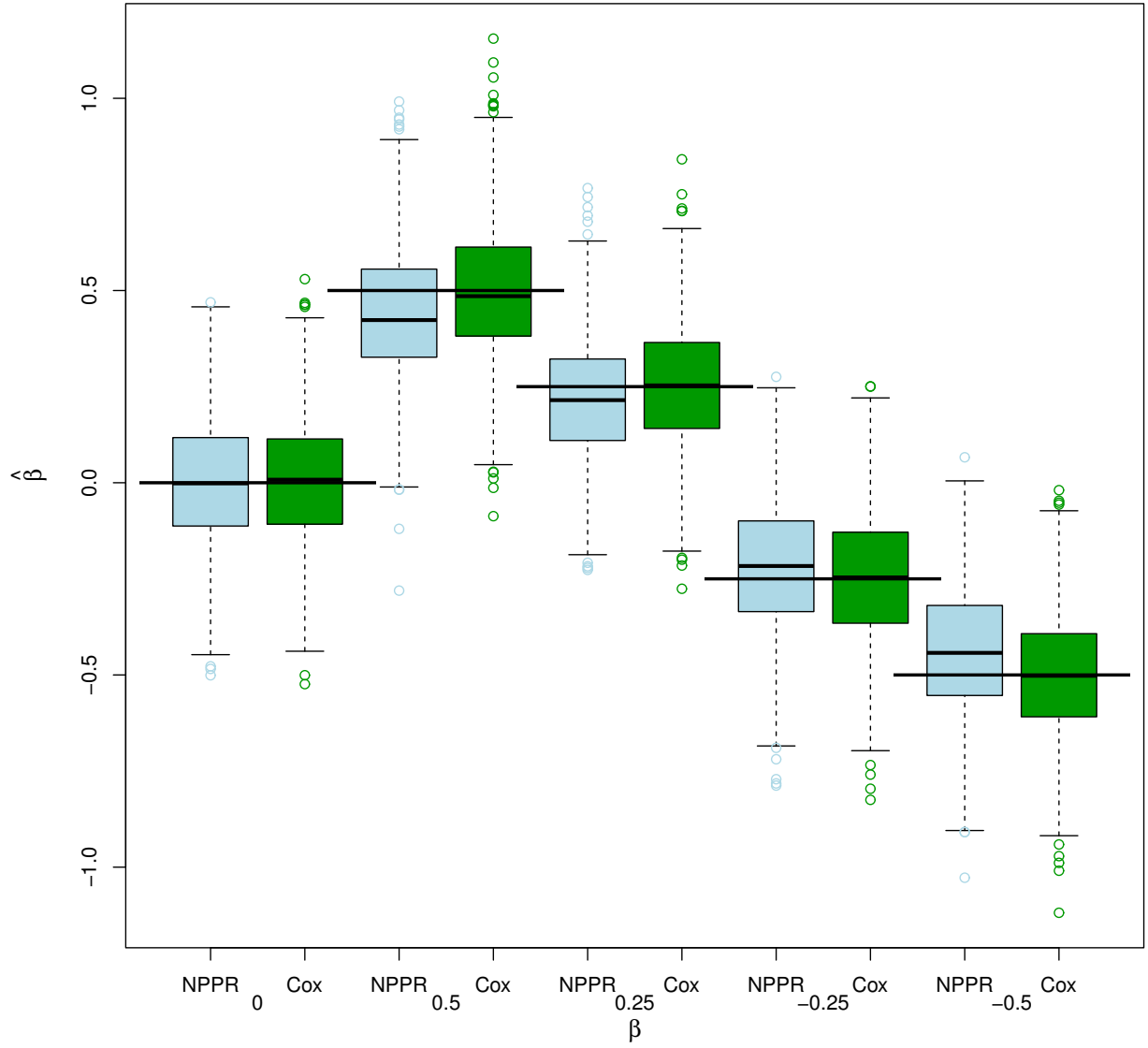


Figure J: Boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR) and $-\log(\widehat{HR}_{Cox})$ (Cox) for all different true β if the Weibull PH model is the underlying model, each in case of 500 participants and 70% censoring rate. Horizontal bars indicate the true underlying effect.

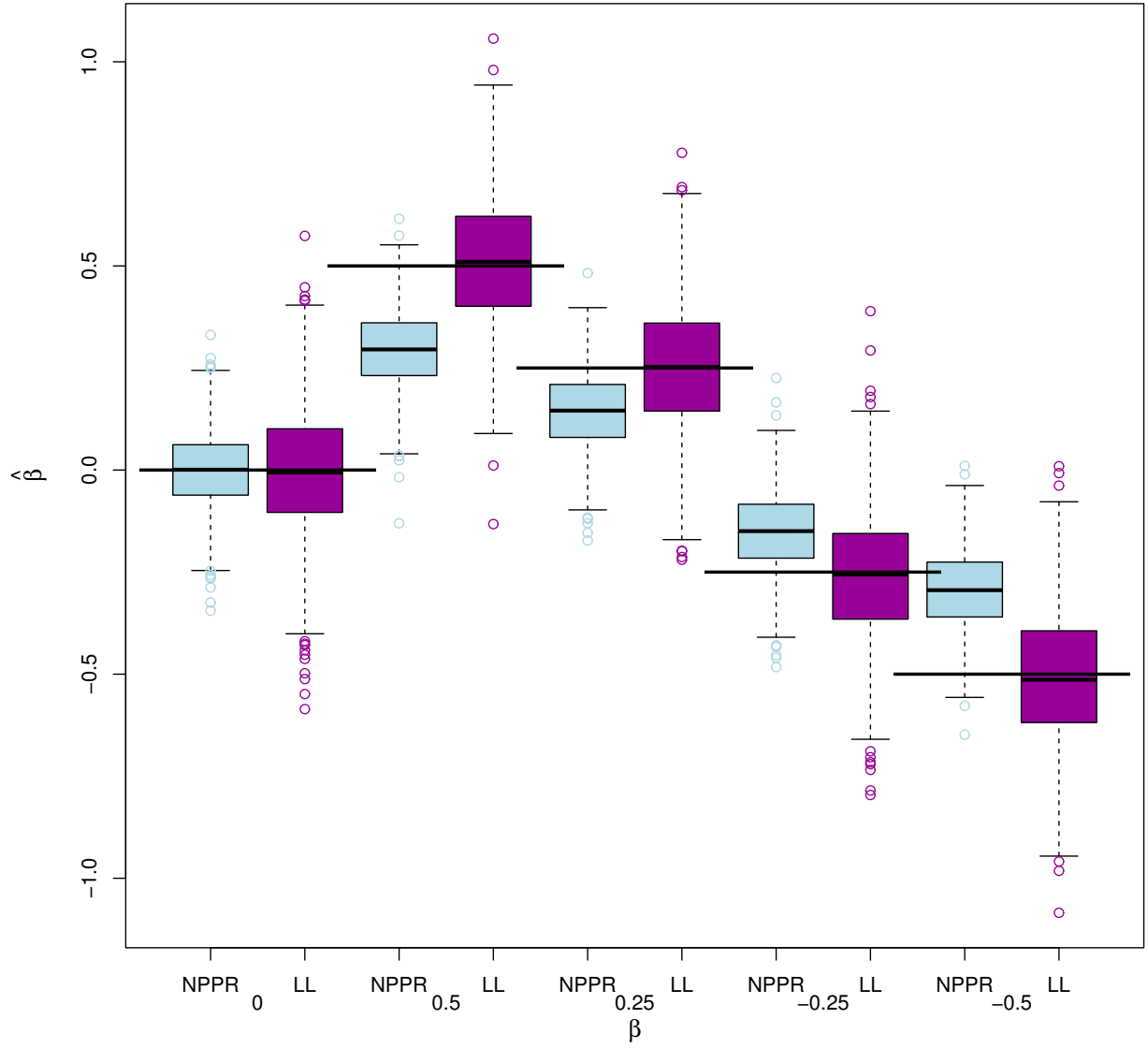


Figure K: Boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR) and $-\log(\widehat{OR}_{LL})$ (LL) for all different true β if the log-logistic PO model is the underlying model, each in case of 500 participants and 30% censoring rate. Horizontal bars indicate the true underlying effect.

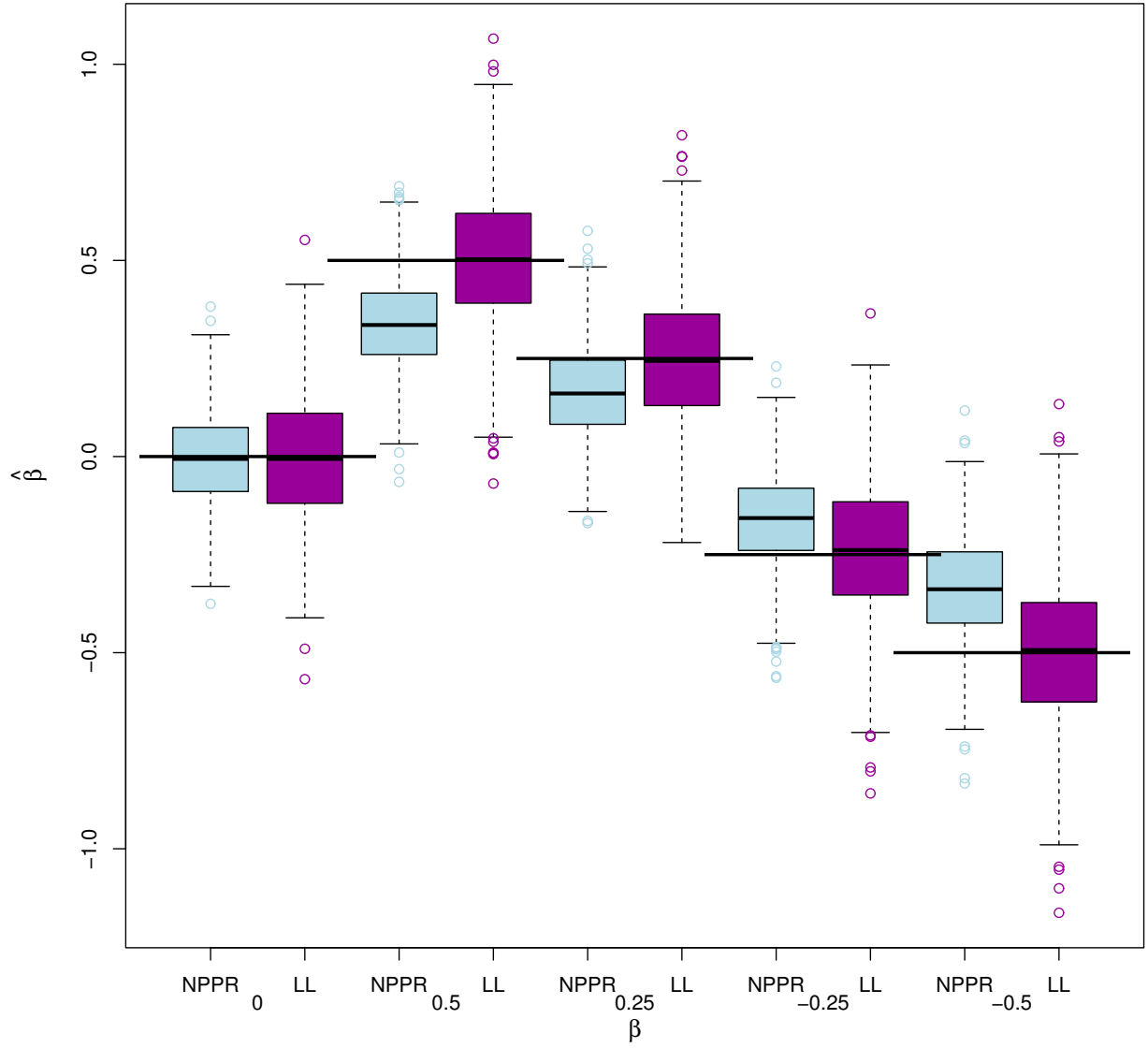


Figure L: Boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR) and $-\log(\widehat{OR}_{LL})$ (LL) for all different true β if the log-logistic PO model is the underlying model, each in case of 500 participants and 50% censoring rate. Horizontal bars indicate the true underlying effect.

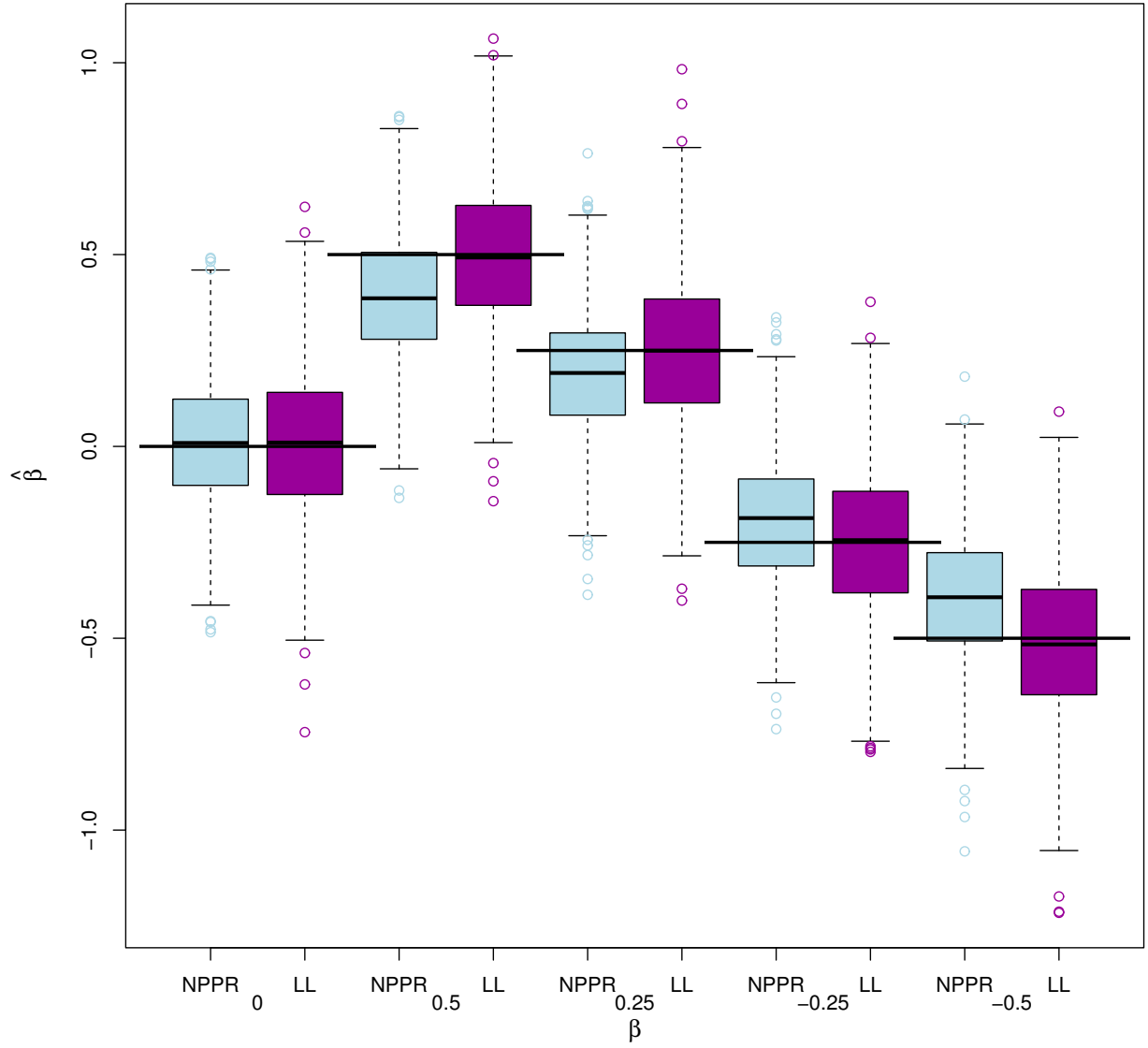


Figure M: Boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR) and $-\log(\widehat{OR}_{LL})$ (LL) for all different true β if the log-logistic PO model is the underlying model, each in case of 500 participants and 70% censoring rate. Horizontal bars indicate the true underlying effect.

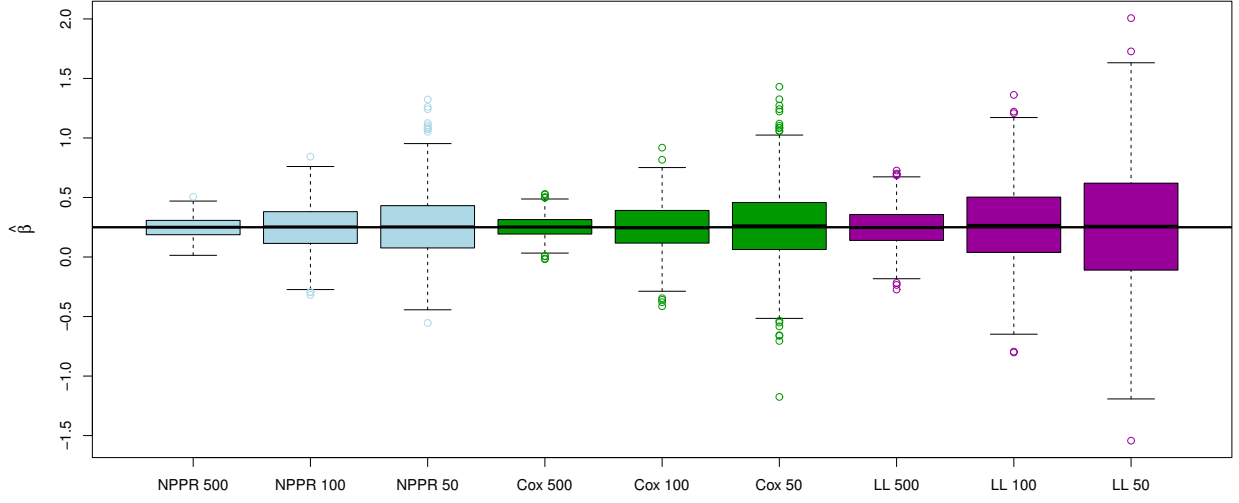


Figure N: Comparison of boxplots for the estimated $\hat{\beta}_{NPPR}$ (NPPR), $-\log(\widehat{HR}_{Cox})$ (Cox) and $-\log(\widehat{OR}_{LL})$ (LL) for different numbers of participants if $\beta = 0.25$, no censoring is present and the respective model assumptions are fulfilled. The horizontal bar indicate the true underlying effect.

3.3.2 Bias and MSE

Table B: Bias and MSE of $\hat{\beta}_{NPPR}$ (NPPR), $-\log(\widehat{HR}_{Cox})$ (Cox) and $-\log(\widehat{OR}_{LL})$ (LL PO) without censoring.

Underlying model	Effect	Participants	Bias			MSE		
			NPPR	Cox	LL	NPPR	Cox	LL
PPR	0.00	500	0.002	0.002	0.004	0.006	0.008	0.021
PPR	0.00	100	-0.005	-0.003	-0.008	0.030	0.045	0.105
PPR	0.00	50	-0.009	-0.008	-0.015	0.056	0.087	0.207
PPR	0.50	500	0.006	0.606	0.509	0.009	0.378	0.281
PPR	0.50	100	0.000	0.591	0.513	0.051	0.405	0.398
PPR	0.50	50	-0.001	0.585	0.532	0.093	0.454	0.549
PPR	0.25	500	-0.001	0.406	0.249	0.007	0.173	0.083
PPR	0.25	100	-0.001	0.400	0.255	0.036	0.204	0.178
PPR	0.25	50	0.014	0.409	0.297	0.072	0.259	0.329
PPR	-0.25	500	0.003	-0.405	-0.247	0.008	0.172	0.083
PPR	-0.25	100	-0.014	-0.416	-0.282	0.039	0.219	0.199
PPR	-0.25	50	-0.002	-0.396	-0.273	0.071	0.245	0.308
PPR	-0.50	500	-0.004	-0.609	-0.509	0.009	0.380	0.282
PPR	-0.50	100	0.006	-0.590	-0.509	0.045	0.403	0.380
PPR	-0.50	50	0.016	-0.580	-0.506	0.088	0.441	0.504
Weibull PH	0.00	500	-0.005	-0.006	-	0.006	0.008	-
Weibull PH	0.00	100	-0.008	-0.003	-	0.033	0.045	-
Weibull PH	0.00	50	0.001	-0.004	-	0.064	0.095	-
Weibull PH	0.50	500	-0.164	-0.001	-	0.034	0.009	-
Weibull PH	0.50	100	-0.173	0.003	-	0.064	0.045	-
Weibull PH	0.50	50	-0.171	0.020	-	0.097	0.104	-
Weibull PH	0.25	500	-0.084	0.004	-	0.014	0.008	-
Weibull PH	0.25	100	-0.085	0.003	-	0.037	0.042	-
Weibull PH	0.25	50	-0.092	0.009	-	0.074	0.096	-
Weibull PH	-0.25	500	0.083	0.000	-	0.013	0.007	-
Weibull PH	-0.25	100	0.079	-0.003	-	0.039	0.042	-
Weibull PH	-0.25	50	0.079	-0.016	-	0.071	0.098	-
Weibull PH	-0.50	500	0.161	-0.003	-	0.033	0.008	-
Weibull PH	-0.50	100	0.167	-0.008	-	0.060	0.042	-
Weibull PH	-0.50	50	0.172	-0.011	-	0.094	0.097	-
LL PO	0.00	500	0.004	-	0.006	0.007	-	0.024
LL PO	0.00	100	0.006	-	0.011	0.032	-	0.127
LL PO	0.00	50	0.004	-	0.013	0.059	-	0.251
LL PO	0.50	500	-0.254	-	-0.004	0.071	-	0.022
LL PO	0.50	100	-0.251	-	0.016	0.096	-	0.133
LL PO	0.50	50	-0.255	-	0.025	0.126	-	0.261
LL PO	0.25	500	-0.127	-	-0.004	0.023	-	0.025
LL PO	0.25	100	-0.117	-	0.021	0.045	-	0.125
LL PO	0.25	50	-0.129	-	0.007	0.080	-	0.272
LL PO	-0.25	500	0.127	-	0.001	0.022	-	0.023
LL PO	-0.25	100	0.137	-	0.019	0.050	-	0.124
LL PO	-0.25	50	0.127	-	-0.009	0.078	-	0.261
LL PO	-0.50	500	0.249	-	-0.003	0.068	-	0.025
LL PO	-0.50	100	0.255	-	0.001	0.096	-	0.127
LL PO	-0.50	50	0.259	-	-0.015	0.131	-	0.269

3.3.3 Coverage

Table C: Coverage (in %) if the PPR is the true underlying model.

Effect	Censoring(%)	Participants	NPPR	PPR	Cox	LL
0.00	30	500	95.3	95.5	94.3	100.0
0.00	30	100	94.7	96.8	93.9	99.9
0.00	30	50	95.0	98.5	94.6	99.9
0.00	50	500	93.9	96.0	94.8	100.0
0.00	50	100	95.6	95.5	94.9	99.7
0.00	50	50	97.5	93.5	94.5	99.6
0.00	70	500	94.2	97.8	94.0	99.8
0.00	70	100	95.9	96.9	95.6	99.6
0.00	70	50	97.8	94.8	96.3	96.7
0.50	30	500	94.6	92.4	1.5	99.4
0.50	30	100	94.6	89.4	58.2	99.3
0.50	30	50	96.1	91.1	81.8	98.6
0.50	50	500	95.5	92.8	18.6	97.5
0.50	50	100	96.5	92.9	79.4	98.5
0.50	50	50	96.5	93.9	90.2	98.4
0.50	70	500	94.3	95.1	86.7	94.4
0.50	70	100	96.4	94.1	93.7	95.8
0.50	70	50	97.6	91.0	96.9	95.3
0.25	30	500	93.9	94.1	22.1	100
0.25	30	100	95.8	93.5	79.0	99.8
0.25	30	50	95.7	97.6	91.0	99.7
0.25	50	500	93.7	94.3	63.2	100
0.25	50	100	94.7	94.6	90.9	99.5
0.25	50	50	96.3	94.0	92.2	99.3
0.25	70	500	94.2	96.2	94.0	99.2
0.25	70	100	95.6	96.2	95.4	97.7
0.25	70	50	97.6	92.0	95.2	97.4
-0.25	30	500	94.6	93.6	20.9	99.9
-0.25	30	100	94.1	97.6	81.9	99.6
-0.25	30	50	95.5	97.6	89.3	100.0
-0.25	50	500	94.3	96.4	72.1	100.0
-0.25	50	100	95.1	97.7	91.8	99.9
-0.25	50	50	96.1	97.2	94.9	99.7
-0.25	70	500	93.7	96.5	92.6	100.0
-0.25	70	100	96.7	96.0	94.0	99.5
-0.25	70	50	98.2	95.6	96.2	98.5
-0.50	30	500	95.7	94.5	0.8	99.2
-0.50	30	100	94.5	95.5	58.6	99.7
-0.50	30	50	95.6	98.5	82.5	99.9
-0.50	50	500	94.1	95.7	16.4	99.9
-0.50	50	100	96.6	98.5	82.0	99.7
-0.50	50	50	96.7	98.4	89.8	99.9
-0.50	70	500	94.3	97.1	90.0	100.0
-0.50	70	100	96.1	97.7	95.8	99.1
-0.50	70	50	97.5	97.3	96.6	98.4

Table D: Coverage (in %) if the Weibull PH model is the true underlying model.

Effect	Censoring(%)	Participants	NPPR	Cox
0.00	30	500	94.4	95.8
0.00	30	100	94.4	94.4
0.00	30	50	95.2	94.1
0.00	50	500	94.7	94.1
0.00	50	100	95.3	96.0
0.00	50	50	96.3	96.5
0.00	70	500	94.9	95.6
0.00	70	100	96.3	95.9
0.00	70	50	98.0	96.3
0.50	30	500	75.6	93.8
0.50	30	100	90.5	95.2
0.50	30	50	94.9	95.8
0.50	50	500	88.7	95.6
0.50	50	100	94.8	94.2
0.50	50	50	96.3	95.7
0.50	70	500	92.4	94.2
0.50	70	100	95.5	95.9
0.50	70	50	97.6	97.0
0.25	30	500	89.9	95.0
0.25	30	100	95.1	95.6
0.25	30	50	96.1	95.3
0.25	50	500	93.9	95.1
0.25	50	100	94.7	94.9
0.25	50	50	94.5	93.9
0.25	70	500	94.7	96.1
0.25	70	100	95.3	94.5
0.25	70	50	97.9	95.8
-0.25	30	500	89.3	94.9
-0.25	30	100	94.9	94.0
-0.25	30	50	95.2	94.8
-0.25	50	500	92.8	95.9
-0.25	50	100	93.5	94.3
-0.25	50	50	96.8	96.2
-0.25	70	500	94.0	94.8
-0.25	70	100	96.7	95.1
-0.25	70	50	97.1	95.9
-0.50	30	500	78.6.9	95.4
-0.50	30	100	90.6	94.6
-0.50	30	50	94.6	95.1
-0.50	50	500	87.8	95.3
-0.50	50	100	93.7	94.5
-0.50	50	50	95.0	96.2
-0.50	70	500	92.7	95.5
-0.50	70	100	96.3	97.1
-0.50	70	50	97.3	95.8

Table E: Coverage (in %) if the log-logistic PO model is the true underlying model.

Effect	Censoring(%)	Participants	NPPR	LL
0.00	30	500	93.9	96.1
0.00	30	100	96.2	97.3
0.00	30	50	96.4	96.4
0.00	50	500	95.8	96.8
0.00	50	100	95.3	95.5
0.00	50	50	97.3	96.0
0.00	70	500	95.8	96.7
0.00	70	100	96.0	96.0
0.00	70	50	97.9	94.4
0.50	30	500	44.5	96.2
0.50	30	100	86.5	96.9
0.50	30	50	89.3	95.4
0.50	50	500	74.8	96.0
0.50	50	100	92.2	97.0
0.50	50	50	93.3	95.4
0.50	70	500	92.7	97.4
0.50	70	100	94.3	95.1
0.50	70	50	96.5	92.5
0.25	30	500	82.3	96.4
0.25	30	100	91.3	96.2
0.25	30	50	95.7	96.2
0.25	50	500	89.0	95.1
0.25	50	100	95.0	96.8
0.25	50	50	95.4	95.4
0.25	70	500	93.6	96.9
0.25	70	100	95.6	96.3
0.25	70	50	98.0	93.9
-0.25	30	500	82.2	96.3
-0.25	30	100	93.0	96.0
-0.25	30	50	94.5	95.3
-0.25	50	500	88.9	97.0
-0.25	50	100	95.7	97.4
-0.25	50	50	96.2	96.2
-0.25	70	500	94.7	96.6
-0.25	70	100	96.0	96.6
-0.25	70	50	97.3	94.7
-0.50	30	500	43.7	96.3
-0.50	30	100	81.6	95.0
-0.50	30	50	89.3	94.7
-0.50	50	500	70.9	95.8
-0.50	50	100	92.5	96.7
-0.50	50	50	94.4	95.4
-0.50	70	500	90.5	97.5
-0.50	70	100	95.3	97.9
-0.50	70	50	97.3	94.0

Table F: Coverage of $\hat{\beta}_{NPPR}$ (NPPR), $-\log(\widehat{HR}_{Cox})$ (Cox's PH model) and $-\log(\widehat{OR}_{LL})$ (log-logistic PO model) if there is no censoring.

Underlying model	Effect	Participants	NPPR	Cox	LL
PPR	0	500	95.2	94.4	99.9
PPR	0	100	94.9	93.4	99.9
PPR	0	50	97.4	95.4	99.7
PPR	0.5	500	95.5	0.0	98.8
PPR	0.5	100	93.4	30.5	99.4
PPR	0.5	50	95.3	65.3	99.2
PPR	0.25	500	95.1	0.3	99.7
PPR	0.25	100	95.4	57.9	99.6
PPR	0.25	50	96.0	78.3	99.5
PPR	-0.25	500	94.9	0.4	99.7
PPR	-0.25	100	94.1	55.2	99.9
PPR	-0.25	50	96.3	80.1	99.9
PPR	-0.5	500	95.8	0.0	98.7
PPR	-0.5	100	95.4	30.6	99.6
PPR	-0.5	50	95.7	65.5	100.0
Weibull PH	0	500	95.5	94.9	-
Weibull PH	0	100	94.3	95.0	-
Weibull PH	0	50	94.9	94.6	-
Weibull PH	0.5	500	49.5	95.1	-
Weibull PH	0.5	100	84.0	95.2	-
Weibull PH	0.5	50	91.1	94.4	-
Weibull PH	0.25	500	80.6	95.3	-
Weibull PH	0.25	100	94.0	95.7	-
Weibull PH	0.25	50	93.8	94.3	-
Weibull PH	-0.25	500	84.3	95.3	-
Weibull PH	-0.25	100	93.1	95.4	-
Weibull PH	-0.25	50	94.6	94.7	-
Weibull PH	-0.5	500	50.7	94.5	-
Weibull PH	-0.5	100	86.3	97.1	-
Weibull PH	-0.5	50	91.2	95.4	-
LL PO	0	500	94.6	-	96.3
LL PO	0	100	95.4	-	95.3
LL PO	0	50	96.1	-	96.0
LL PO	0.5	500	10.7	-	97.5
LL PO	0.5	100	71.0	-	95.8
LL PO	0.5	50	83.4	-	95.9
LL PO	0.25	500	64.2	-	96.0
LL PO	0.25	100	89.9	-	96.7
LL PO	0.25	50	93.1	-	95.3
LL PO	-0.25	500	64.7	-	96.0
LL PO	-0.25	100	88.9	-	95.7
LL PO	-0.25	50	93.0	-	95.4
LL PO	-0.5	500	11.8	-	95.5
LL PO	-0.5	100	71.5	-	96.5
LL PO	-0.5	50	82.2	-	96.4

3.3.4 Numerical Robustness

Table G: Numerical Robustness in 1000 studies if the PPR model is the true underlying model.

Effect	Censoring(%)	Participants	NPPR	PPR	Cox	LL
0.00	30	500	1000	993	1000	1000
0.00	30	100	1000	835	1000	1000
0.00	30	50	1000	748	1000	1000
0.00	50	500	1000	1000	1000	1000
0.00	50	100	1000	997	1000	1000
0.00	50	50	1000	979	1000	1000
0.00	70	500	1000	1000	1000	1000
0.00	70	100	1000	999	1000	1000
0.00	70	50	997	991	999	999
0.50	30	500	1000	984	1000	1000
0.50	30	100	1000	866	1000	1000
0.50	30	50	1000	673	1000	999
0.50	50	500	1000	997	1000	1000
0.50	50	100	1000	959	1000	1000
0.50	50	50	1000	882	1000	999
0.50	70	500	1000	1000	1000	1000
0.50	70	100	1000	997	1000	1000
0.50	70	50	991	983	995	994
0.25	30	500	1000	989	1000	1000
0.25	30	100	1000	837	1000	1000
0.25	30	50	1000	761	1000	1000
0.25	50	500	1000	999	1000	1000
0.25	50	100	1000	968	1000	1000
0.25	50	50	1000	938	1000	999
0.25	70	500	1000	1000	1000	1000
0.25	70	100	1000	1000	1000	1000
0.25	70	50	998	992	999	999
-0.25	30	500	1000	993	1000	1000
-0.25	30	100	1000	807	1000	1000
-0.25	30	50	1000	734	1000	1000
-0.25	50	500	1000	1000	1000	1000
-0.25	50	100	1000	982	1000	1000
-0.25	50	50	1000	949	1000	1000
-0.25	70	500	1000	1000	1000	1000
-0.25	70	100	1000	1000	1000	1000
-0.25	70	50	993	985	998	997
-0.50	30	500	1000	981	1000	1000
-0.50	30	100	1000	757	1000	1000
-0.50	30	50	1000	701	1000	1000
-0.50	50	500	1000	997	1000	1000
-0.50	50	100	1000	937	1000	1000
-0.50	50	50	1000	876	1000	1000
-0.50	70	500	1000	1000	1000	1000
-0.50	70	100	1000	999	1000	1000
-0.50	70	50	988	989	998	998

Table H: Numerical Robustness in 1000 studies if the Weibull PH model is the true underlying model.

Effect	Censoring(%)	Participants	NPPR	Cox
0.00	30	500	1000	1000
0.00	30	100	1000	1000
0.00	30	50	1000	1000
0.00	50	500	1000	1000
0.00	50	100	1000	1000
0.00	50	50	1000	1000
0.00	70	500	1000	1000
0.00	70	100	1000	1000
0.00	70	50	994	1000
0.50	30	500	1000	1000
0.50	30	100	1000	1000
0.50	30	50	1000	1000
0.50	50	500	1000	1000
0.50	50	100	1000	1000
0.50	50	50	1000	1000
0.50	70	500	1000	1000
0.50	70	100	1000	1000
0.50	70	50	991	999
0.25	30	500	1000	1000
0.25	30	100	1000	1000
0.25	30	50	1000	1000
0.25	50	500	1000	1000
0.25	50	100	1000	1000
0.25	50	50	1000	1000
0.25	70	500	1000	1000
0.25	70	100	1000	1000
0.25	70	50	993	997
-0.25	30	500	1000	1000
-0.25	30	100	1000	1000
-0.25	30	50	1000	1000
-0.25	50	500	1000	1000
-0.25	50	100	1000	1000
-0.25	50	50	1000	1000
-0.25	70	500	1000	1000
-0.25	70	100	1000	1000
-0.25	70	50	993	998
-0.50	30	500	1000	1000
-0.50	30	100	1000	1000
-0.50	30	50	1000	1000
-0.50	50	500	1000	1000
-0.50	50	100	1000	1000
-0.50	50	50	1000	999
-0.50	70	500	1000	1000
-0.50	70	100	1000	1000
-0.50	70	50	989	998

Table I: Numerical Robustness in 1000 studies if the log-logistic PO model is the true underlying model.

Effect	Censoring(%)	Participants	NPPR	LL
0.00	30	500	1000	1000
0.00	30	100	1000	1000
0.00	30	50	1000	1000
0.00	50	500	1000	1000
0.00	50	100	1000	1000
0.00	50	50	1000	1000
0.00	70	500	1000	1000
0.00	70	100	1000	1000
0.00	70	50	996	1000
0.50	30	500	1000	1000
0.50	30	100	1000	1000
0.50	30	50	1000	1000
0.50	50	500	1000	1000
0.50	50	100	1000	1000
0.50	50	50	1000	1000
0.50	70	500	1000	1000
0.50	70	100	1000	1000
0.50	70	50	987	995
0.25	30	500	1000	1000
0.25	30	100	1000	1000
0.25	30	50	1000	1000
0.25	50	500	1000	1000
0.25	50	100	1000	1000
0.25	50	50	1000	1000
0.25	70	500	1000	1000
0.25	70	100	1000	1000
0.25	70	50	993	998
-0.25	30	500	1000	1000
-0.25	30	100	1000	1000
-0.25	30	50	1000	1000
-0.25	50	500	1000	1000
-0.25	50	100	1000	1000
-0.25	50	50	1000	1000
-0.25	70	500	1000	1000
-0.25	70	100	1000	1000
-0.25	70	50	992	998
-0.50	30	500	1000	1000
-0.50	30	100	1000	1000
-0.50	30	50	1000	1000
-0.50	50	500	1000	1000
-0.50	50	100	1000	1000
-0.50	50	50	1000	1000
-0.50	70	500	1000	1000
-0.50	70	100	1000	1000
-0.50	70	50	991	998

Table J: Numerical robustness of $\hat{\beta}_{NPPR}$ (NPPR), $-\log(\widehat{HR}_{Cox})$ (Cox's PH model) and $-\log(\widehat{OR}_{LL})$ (log-logistic PO model) if there is no censoring.

Underlying model	Effect	Participants	NPPR	Cox	LL
PPR	0	500	1000	1000	1000
PPR	0	100	1000	1000	1000
PPR	0	50	1000	1000	1000
PPR	0.5	500	1000	1000	1000
PPR	0.5	100	1000	1000	1000
PPR	0.5	50	1000	1000	999
PPR	0.25	500	1000	1000	1000
PPR	0.25	100	1000	1000	1000
PPR	0.25	50	1000	1000	1000
PPR	-0.25	500	1000	1000	1000
PPR	-0.25	100	1000	1000	1000
PPR	-0.25	50	1000	1000	1000
PPR	-0.5	500	1000	1000	1000
PPR	-0.5	100	1000	1000	1000
PPR	-0.5	50	1000	1000	1000
Weibull PH	0	500	1000	1000	-
Weibull PH	0	100	1000	1000	-
Weibull PH	0	50	1000	1000	-
Weibull PH	0.5	500	1000	1000	-
Weibull PH	0.5	100	1000	1000	-
Weibull PH	0.5	50	1000	1000	-
Weibull PH	0.25	500	1000	1000	-
Weibull PH	0.25	100	1000	1000	-
Weibull PH	0.25	50	1000	1000	-
Weibull PH	-0.25	500	1000	1000	-
Weibull PH	-0.25	100	1000	1000	-
Weibull PH	-0.25	50	1000	1000	-
Weibull PH	-0.5	500	1000	1000	-
Weibull PH	-0.5	100	1000	1000	-
Weibull PH	-0.5	50	1000	1000	-
LL PO	0	500	1000	-	1000
LL PO	0	100	1000	-	1000
LL PO	0	50	1000	-	1000
LL PO	0.5	500	1000	-	1000
LL PO	0.5	100	1000	-	1000
LL PO	0.5	50	1000	-	1000
LL PO	0.25	500	1000	-	1000
LL PO	0.25	100	1000	-	1000
LL PO	0.25	50	1000	-	1000
LL PO	-0.25	500	1000	-	1000
LL PO	-0.25	100	1000	-	1000
LL PO	-0.25	50	1000	-	1000
LL PO	-0.5	500	1000	-	1000
LL PO	-0.5	100	1000	-	1000
LL PO	-0.5	50	1000	-	1000

4 Case study: DAPA-HF trial

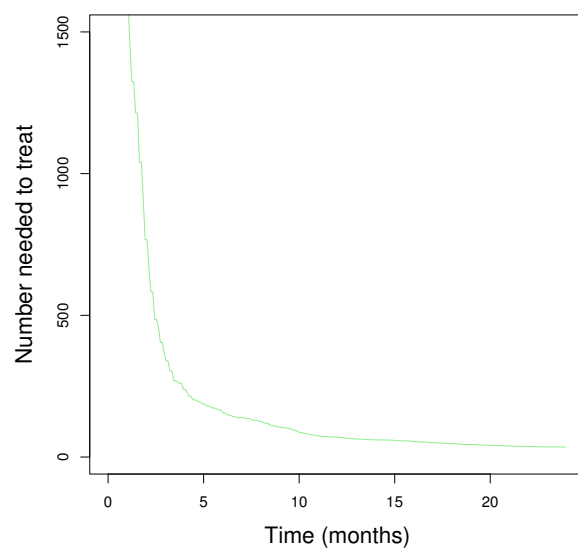


Figure O: Estimated number needed to treat over time for the DAPA-HF study.

3.2 Identification of changes in gene expression

This article discusses a shift in the underlying research question in model-based alert identification for one continuous covariate. It introduces an approach with the objective of identifying the entire covariate region over which the response changes. This is achieved by fitting a parametric model and conducting a hypothesis test based on its first derivative, such that rejecting the null hypothesis means detecting a significant change in response. The test decision is obtained by estimating a simultaneous confidence band for the first derivative using a two-step bootstrap procedure. In comparison with existing methodologies, the proposed method extends the analysis beyond individual points. Its performance is investigated by means of a simulation study and a variety of applications to mouse gene expression data from a study investigating the effect of a Western diet on the progression of non-alcoholic fatty liver disease are demonstrated.

Authorship:	First author
Coauthors:	Kathrin Möllenhoff
Contribution statement:	My contributions to this project included developing the method and implementing it in R, including programming, executing the simulation studies and case study, preparing figures and tables, and writing the first draft of the manuscript. The initial idea was developed by me and Kathrin Möllenhoff. Kathrin Möllenhoff also supported the development of the method. All coauthors discussed the contents and structure and revised the manuscript.
Status:	Accepted (publication pending)
Journal:	Biometrical Journal (preprint of unrevised version available on arXiv)
DOI:	arXiv.2407.14630
Note:	Here, the revised version is presented

Identification of changes in gene expression

L. Ameis¹ and K. Möllenhoff^{1*}

¹Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine,
University of Cologne

Abstract

Abstract. Evaluating the change in gene expression is a common goal in many research areas, such as in toxicological studies, which are particularly important in pre-clinical research. In practice, the analysis is often based on multiple t-tests evaluated at the observed time points of the experiment, which limits the accuracy of determining the precise time at which the gene changes in expression. If a parametric approach is chosen, the analysis is often restricted to identifying the onset of an effect, but not its length. In this paper, we propose a parametric method to identify the time frame where the gene expression significantly changes. This is achieved by fitting a parametric model and constructing a confidence band for its first derivative. The confidence band is derived by a two-step bootstrap approach. It is summarised in terms of a hypothesis test, such that rejecting the null hypothesis means detecting a significant change in gene expression. Furthermore, a method for calculating confidence intervals for time points of interest (e.g. the beginning of significant change) is developed. We demonstrate the validity of our approach through a simulation study and present a variety of different applications to mouse gene expression data from a study investigating the effect of a Western diet on the progression of non-alcoholic fatty liver disease.

1 Introduction

Understanding the mechanisms underlying disease responses is a key objective of toxicological studies, which are particularly important in preclinical research. A powerful tool in molecular biology for this purpose is RNA-sequencing, based on the measurement of transcriptomes (Wang, Z. and Gerstein, M. and Snyder, M., 2009), which is integral to our understanding of genomic function (Stark, R. and Grzelak, M. and Hadfield, J., 2019). It measures the expressions of many genes simultaneously and is extensively used in drug discovery and development (Khatoon, Z. and Figler, B. and Zhang, H. and Cheng, F., 2014). One common objective in the field of RNA-sequencing is to identify differentially expressed genes. Possible examples include dose-gene expression data (Krug, A.K. et al., 2013) or time-gene expression data (Ghallab, A. et al., 2021).

Anders, S. and Huber, W. (2010) introduced an approach to analysing gene count data using the negative binomial distribution, that is employed in the well-known DESeq R package. Its successor – the DESeq2 packages (Love, M.I. and Huber, W. and Anders, S., 2014) – has been cited over 35,000 times as of July 2024 (NCBI, 1996). However, as described in the vignette of DESeq2 (Love, M.I. and Huber, W. and Anders, S., 2024), the analysis of the data includes the dose/time as a discrete covariate, which limits the precision of the analysis to the measured dose levels/time points.

One question, where this problem is particularly evident, is the identification of an alert concentration, defined as the lowest measured concentration where the difference between the concentration and the control significantly exceeds a critical level of the relevant response (Delignette-Muller, M.-L. and Forfait, C. and Billoir, E. and Charles, S., 2011) or the lowest concentration with a noticeable effect (Jensen, S.M. and Kluxen, F.M. and Ritz,

C., 2019). Traditionally, the analysis is based on applying multiple t -tests or – as recommended – Dunnett tests (Hothorn, L.A., 2014), focusing on the measured dose levels. Consequently, with these tests it is not possible to identify a dosage between those levels. In order to address this issue, Kappenberg, F. and Grinberg, M. and Jiang, X. and Kopp-Schneider, A. and Hengstler, J.G. and Rahnenführer, J. (2021) proposed a model-based approach assuming a monotonic relationship between dose and response. This method allows for a continuous identification of the lowest concentration at which the response significantly exceeds the response for the control by a given threshold. This method was further developed by Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022) through the introduction of a procedure that is independent of the assumption of monotonicity and replaces an asymptotic calculation of the variance by a parametric bootstrap.

When analysing time-response data, a similar problem is given by the identification of the time frame of significant changes in gene expression. For example, a change in gene expression indicates deregulation and thus the presence of a biological process. In a two-group situation, a difference in time periods may indicate a slowing or acceleration of an underlying process. In addition, a Gene Ontology (GO) enrichment analysis of genes that undergo a change in expression over a given time period allows conclusions to be drawn about the order of processes occurring at the organism level, which can be used in particular to assess disease progression. In Ghallab, A. et al. (2021) so called rest-and-jump genes (RJG) were observed. These are genes that showed delayed deregulation, either reaching a plateau afterwards or continuing to deregulate. In this context, it is of interest to consider the entire time frame during which the gene expression changes, rather than focusing only on the onset of an effect.

Based on this idea, in this paper we propose a model-based approach that allows the identification of the entire time frame within which the change in gene expression is considered significant. This is accomplished by formulating a hypothesis test considering a test statistic based on the absolute value of the first derivative of a selected continuous parametric model or a B-spline fit with regard to a threshold $\lambda \geq 0$. Rejecting the null hypothesis indicates the presence of a slope, that is significantly larger than λ and thereby a significant change in gene expression. Finally, the test decision is based on the construction of a lower simultaneous confidence band for the first derivative.

The proposed approach focuses on biologically significant changes in expression and the specific expression pattern of single or multiple genes. It therefore allows a more explicit interpretation of the results than, for example, classification by clustering (Androulakis, I.P. and Yang, E. and Almon, R.R., 2007). Especially in the case of toxicological data, recent publications suggest the use of parametric approaches (Kappenberg, F. and Duda, J.C. and Schürmeyer, L. and Gül, O. and Brecklinghaus, T. and Hengstler, J.G. and Schorning, K. and Rahnenführer, J., 2023). The first derivative of a parametric model fit to the response data provides an understandable and easy to implement translation from the gene count scale to a measure of changes in expression. Similarly, the first derivative of a model fit utilizing B-splines is itself a linear combination of a B-spline basis (Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D., 2022). Both approaches are therefore well suited to analysing expression change explicitly. The construction of simultaneous confidence bands presented here is specifically tailored to the first derivative, in contrast to established approaches that are fit to the response data (Neumann, M.H. and Polzehl, J., 1998; Wang, L. and Yang, L., 2010).

This paper is structured as follows: First, we introduce the method to estimate the time frame of significant change in gene expression, based on calculating the first derivative of the estimated time-response model and the estimation of a corresponding lower simultaneous confidence band. Second, the methodology is subjected to a simulation study. Finally, potential applications are illustrated using a data set from the western diet mice study (Ghallab, A. et al., 2021), a toxicological serial-sacrifice study (Dewanji, A., 2005) that examines the impact of a western diet (WD) on the progression of non-alcoholic fatty liver disease (NAFLD).

2 Methodology

The fundamental idea of this method is to initially fit a model to the data and subsequently assess the change in gene expression by considering its first derivative. This is possible since the first derivative describes the slope of the gene expression curve at any time point t , and thus can be regarded as a measure for the change. Once a suitable model has been fitted, the next step is to develop a hypothesis test that tests if and when the first derivative significantly exceeds a chosen threshold, denoted λ . The corresponding time point marks a change in gene expression. The procedure is inspired by the method developed in Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022), adapted for the first derivative.

2.1 Model fit and derivation

We regard a data set with n observations in total, that are conducted at m different time points $t_q, q = 1, \dots, m$. Let n_q denote the number of observations at time point t_q , i.e. $\sum_{q=1}^m n_q = n$. The presented method is designed to accommodate for repeated measurements. However, it can be simplified in case of an independent structure along t . In cases where there is a correlation between measurements at different time points, the same number of measurements, $n_1 = \dots = n_q$, are assumed for the purposes of simplicity. Furthermore, the study period is defined as $T = [t_1, t_m]$ and the study duration as $\bar{t} = t_m - t_1$.

The initial step is to fit a parametric model to the data. We define

$$y_{p,q} = f(t_q, \boldsymbol{\theta}) + \varepsilon_{p,q} \quad p = 1, \dots, n_q, \quad q = 1, \dots, m,$$

where f denotes the selected model. In order to incorporate possible correlations, we assume that the error terms $\boldsymbol{\varepsilon}_p = (\varepsilon_{p,1}, \dots, \varepsilon_{p,m})$ are multivariate normal distributed $\boldsymbol{\varepsilon}_p \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$ and independent for $p = 1, \dots, n_q$. The parameters $\boldsymbol{\theta} \in B \subset \mathbb{R}^r$ are estimated by applying the Maximum likelihood method. The likelihood is given by

$$\mathcal{L}_{\mathbf{y}, \mathbf{t}}(\boldsymbol{\theta}) = \prod_{p=1}^{n_q} \frac{\exp\left(-\frac{1}{2}(\mathbf{y}_p - f(\mathbf{t}, \boldsymbol{\theta}))^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}_p - f(\mathbf{t}, \boldsymbol{\theta}))\right)}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}}$$

with $\mathbf{y}_p = (y_{p,1}, \dots, y_{p,m})$ and $\mathbf{t} = (t_1, \dots, t_m)$. The estimator for the covariance matrix is given by

$$\hat{\boldsymbol{\Sigma}}^2 = \frac{1}{n_q} \sum_{p=1}^{n_q} (\hat{\boldsymbol{\varepsilon}}_p - \hat{\boldsymbol{\varepsilon}})(\hat{\boldsymbol{\varepsilon}}_p - \hat{\boldsymbol{\varepsilon}})^T,$$

whereby the vector $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_m)$ consists of the means over all measurements

$$\hat{\varepsilon}_{p,q} = y_{p,q} - f(t_q, \hat{\boldsymbol{\theta}}) \text{ for } q = 1, \dots, m.$$

In order to guarantee a good fit, the model is selected from a pool of possible models using a selection criteria, as, for example, the AIC (Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D., 2022).

As we are interested in detecting significant changes, the method developed here does not focus on the model f itself, but rather on its first derivative f' . Motivated by the findings in Duda, J.C. and Kappenberg, F. and Rahnenführer, J. (2022), we will focus on applying the method using the 4pLL (also called sigmoid Emax) model, the beta model, the exponential model and the quadratic model, respectively, as parameterized in Bornkamp, B. and Pinheiro, J. and Bretz, F. (2009). For details as well as the different first derivatives we refer to the supplementary material.

Another focus will be on utilizing B-splines for model fitting as a more flexible alternative. As stated in Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D. (2022), a B-spline basis function of degree l is constructed from $(l+1)$ piecewise polynomials of degree l that are joined, such that the basis function is $(l-1)$ -times continuously

differentiable. For the construction, the domain of the covariate, here time, is partitioned into intervals by so-called knots. A full B-spline basis consists of a set of the basis functions set up along these knots. The full model f is a linear combination of the basis. Naturally, since the basis functions are $(l - 1)$ -times continuously differentiable, so is the f . Therefore, B-splines of degree $l \geq 2$ can be used to apply the proposed method. For details on the recursive definition of the B-spline basis functions and their first derivatives, we refer to the supplementary material.

Nevertheless, the method is applicable to all parametric models with a first derivative that can be calculated analytically.

2.2 Hypothesis test for the identification of significant changes

Here, we introduce a hypothesis test in which the rejection of H_0 leads to the assertion of a significant change in gene expression. In the previously introduced parametric framework this means that f' exceeds a threshold $\lambda \geq 0$ at some time point t , which leads to the hypotheses

$$H_0 : \forall t \in T \ |f'(t, \boldsymbol{\theta})| \leq \lambda \text{ vs. } H_1 : \exists t \in T \ |f'(t, \boldsymbol{\theta})| > \lambda. \quad (1)$$

Rejecting H_0 implies the existence of at least one time point at which a change in gene expression is detected that exceeds the critical value λ . It is important to note that λ should be selected beforehand. Should any change in gene expression be of interest, λ can be set to 0. However, if only a larger change may be biologically relevant, λ should be adjusted accordingly in consultation with experts. For instance, regarding the absolute difference of the model fit to $\log_2$ -transformed counts at two time points t_a and t_b , i.e. $|f(t_a, \boldsymbol{\theta}) - f(t_b, \boldsymbol{\theta})|$, the corresponding threshold is often set to $\log_2(1.5)$ (Ghallab, A. et al., 2021; Kappenberg, F. and Grinberg, M. and Jiang, X. and Kopp-Schneider, A. and Hengstler, J.G. and Rahnenführer, J., 2021; Möllenhoff, K. and Schorning, K. and Kappenberg, F., 2022), investigating whether the foldchange between t_a and t_b exceeds 1.5. Based on this, a possible choice would be given by $\lambda = \log_2(1.5) \times t^{-1}$ for the test regarding f' , representing the slope of a straight line that increases by $\log_1(1.5)$ between t_1 to t_m . Of note, $\bar{t}$ is interchangeable with any other time frame, e.g. $(1/2)\bar{t}$. Then the question is whether there is an absolute change in $\log_2(1.5)$ over half of the study period.

Next, let $L^\alpha(t, \hat{\boldsymbol{\theta}})$ denote a $(1 - \alpha)$ lower simultaneous confidence band of $|f'(t, \boldsymbol{\theta})|$, that is

$$\mathbb{P}\left\{\forall t \in T : L^\alpha(t, \hat{\boldsymbol{\theta}}) \leq |f'(t, \boldsymbol{\theta})|\right\} \geq 1 - \alpha.$$

According to Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022), rejecting H_0 in (1) if $L^\alpha(t, \hat{\boldsymbol{\theta}}) > \lambda$ for any $t \in T$, leads to an α -level test, as

$$\begin{aligned} \mathbb{P}_{H_0}\left\{\exists t \in T : L^\alpha(t, \hat{\boldsymbol{\theta}}) > \lambda\right\} &\leq \mathbb{P}_{H_0}\left\{\exists t \in T : L^\alpha(t, \hat{\boldsymbol{\theta}}) > |f'(t, \boldsymbol{\theta})|\right\} \\ &= 1 - \mathbb{P}_{H_0}\left\{\forall t \in T : L^\alpha(t, \hat{\boldsymbol{\theta}}) \leq |f'(t, \boldsymbol{\theta})|\right\} \\ &\leq 1 - (1 - \alpha) = \alpha. \end{aligned}$$

The time period of significant change in gene expression corresponds to the set of all $t \in T$ for which H_0 in (1) can be rejected, which we denote by

$$\tilde{S} := \left\{t : L^\alpha(t, \hat{\boldsymbol{\theta}}) > \lambda\right\}.$$

Figure 1 A depicts an illustrative example of a first derivative of a gene expression curve f' (black) and its lower confidence band (blue). The orange line represents the time period $\tilde{S}$, during which the confidence band exceeds the threshold λ , indicated by the red line.

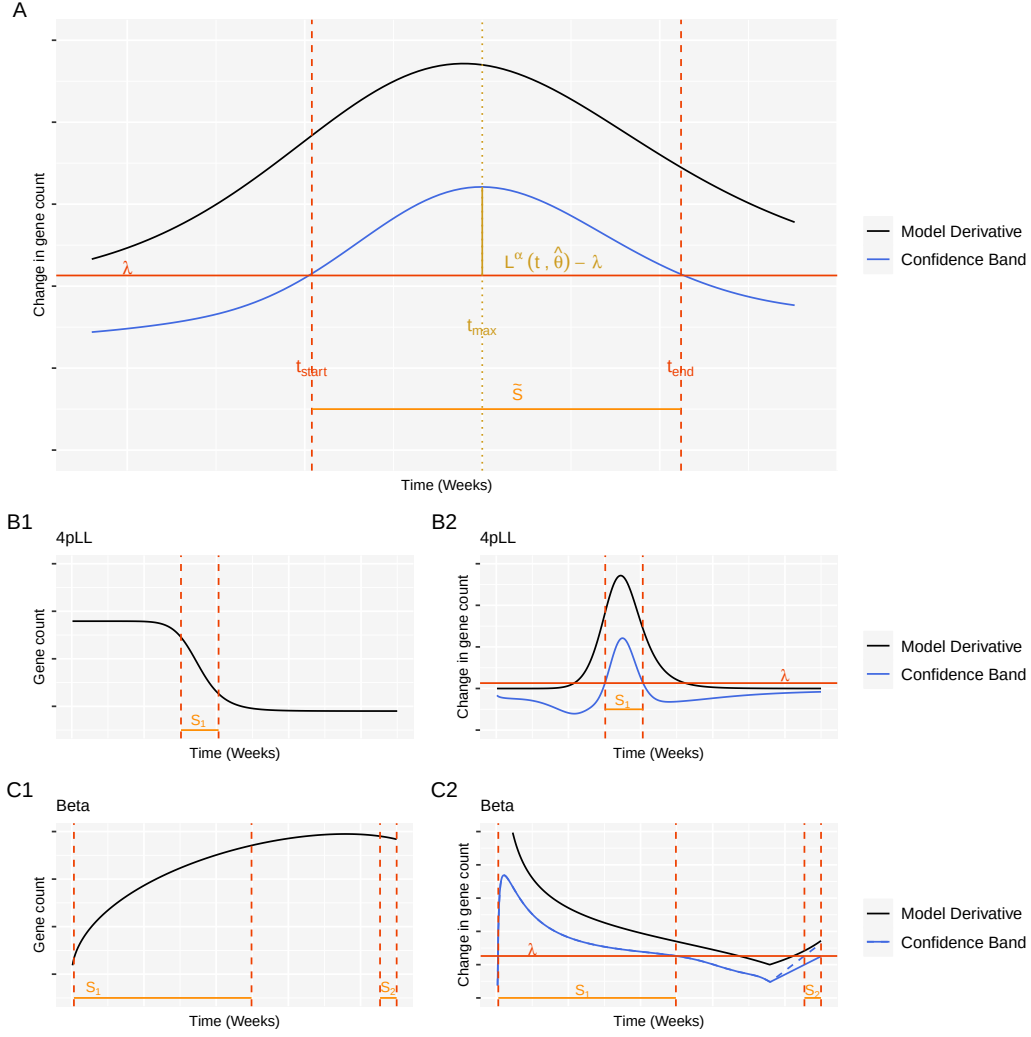


Figure 1: (A) Visualisation of the time period of significant change in gene expression $\tilde{S}$ (orange) with two time points of interest. The red dotted lines indicate the beginning and end of $\tilde{S}$. The yellow line indicates the time point of maximum change in expression. $\tilde{S}$ is partitioned into coherent subsets if the gene expression curve follows a 4pLL model (B) or a beta model (C). B1 and C1 depict the fitted model, B2 and C2 the corresponding first derivative and confidence bands such that H_0 in (1) is rejected.

While we can conclude significant changes in gene expression for all $t \in T$ which fulfill $L^\alpha(t, \hat{\theta}) > \lambda$, some *specific time points* might be of higher interest. An example are the beginning and end of a significant change. Precisely, these are the first time point $t_{\text{start}} \in T$ with $L^\alpha(t_{\text{start}}, \hat{\theta}) > \lambda$ and the last time point $t_{\text{end}} \in T$ with $L^\alpha(t_{\text{end}}, \hat{\theta}) > \lambda$. Figure 1 A indicates both t_{start} and t_{end} by the red vertical lines and thus also the period of significant change $\tilde{S}$. Another example is the time point of maximal change, indicated by the yellow vertical line in Figure 1 A, representing the time point $t_{\text{max}} \in T$ at which $L^\alpha(t, \hat{\theta}) - \lambda$, $t \in \tilde{S}$ assumes its (locally) maximal value.

These time points of interest are not necessarily unique. In case of a non-monotonous function f (e.g. beta model), the confidence band can dip below the threshold λ and then exceed it again after a time period of no significant change. Therefore, we can partition S into ascending coherent subsets $\tilde{S} = (S_1, \dots, S_s)$, meaning that for all $j = 1, \dots, s-1$, it exists a $t' \in T$ between S_j and S_{j+1} such that $L^\alpha(t', \hat{\theta}) < \lambda$ and $\max(S_j) < \min(S_{j+1})$.

We refer to Figure 1 B and C for a visual representation. Figure 1 B depicts a gene expression curve that follows a 4pLL model on the left and its first derivative with a possible confidence band on the right. Here, H_0 is rejected and the time period of significant change in gene expression is coherent. Then $\tilde{S}$ equals S_1 . Figure 1 C depicts a beta model in the left and its first derivative with two possible confidence bands on the right. The solid line represents the scenario in which the decline in gene count in the latter part of the study period is not statistically significant. The dashed line on the other hand represents the scenario in which the aforementioned decline is significant. Here, $\tilde{S}$ splits into the coherent sub-periods S_1 and S_2 . Each S_j , $j = 1, \dots, s$, has a unique set of the aforementioned time points of interest. In the following, we will focus on t_{start} and t_{end} as an example. Therefore, the set of times to search for is given by

$$\left\{ t_{\text{start}}^{(j)} = \min(S_j) : j = 1, \dots, s \right\} \cup \left\{ t_{\text{end}}^{(j)} = \max(S_j) : j = 1, \dots, s \right\}. \quad (2)$$

The precise estimation of these quantities depends on the accuracy of the estimation of $L^\alpha(t, \hat{\theta})$, which is described in the next section.

2.3 Estimating a confidence band of $|f'(t)|$

In order to estimate a lower simultaneous confidence band $L^\alpha(t, \hat{\theta})$ of $|f'(t, \theta)|$ the method developed in Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022) is adjusted to the first derivative and to account for a dependence structure along the time axis. Therefore, we define

$$L^\alpha(t, \hat{\theta}) := |f'(t, \hat{\theta})| - c\hat{\sigma}_{|f'(t, \hat{\theta})|}$$

and estimate the critical value c , such that

$$\mathbb{P} \left\{ \max_{t \in T} \frac{|f'(t, \hat{\theta})| - |f'(t, \theta)|}{\hat{\sigma}_{|f'(t, \hat{\theta})|}} \leq c \right\} = 1 - \alpha.$$

To estimate both $\hat{\sigma}_{|f'(t, \hat{\theta})|}$ and c , a two-step parametric bootstrap procedure inspired by the classic bootstrap-t method (Efron, B. and Tibshirani, R.J., 1994) has been proposed in Algorithm 1 by Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022) and is applied here by substituting the difference for the first derivative. In short, $\hat{\theta}$ and $\hat{\Sigma}$ are estimated from the original data set and used to generate B_1 bootstrap samples. This is done by randomly drawing iid. errors $\varepsilon_p^* = (\varepsilon_{p,1}^*, \dots, \varepsilon_{p,m}^*) \sim N_m(\mathbf{0}, \hat{\Sigma})$ and applying the formula

$$y_{p,q}^* = f(t_q, \hat{\theta}) + \varepsilon_{p,q}^* \text{ for } p = 1, \dots, n_q \text{ and } q = 1, \dots, m.$$

These can be considered the first level of the bootstrap. For each of the generated samples, $\hat{\theta}_l^*$, $l = 1, \dots, B_1$, are estimated and used to generate another B_2 bootstrap samples, which constitute the second level of the bootstrap. Then, $\hat{\sigma}_{|f'(t, \hat{\theta})|}$ is estimated from the first bootstrap level by calculating the empirical variance of the sample $|f'(t, \hat{\theta}_1^*)|, \dots, |f'(t, \hat{\theta}_{B_1}^*)|$. Similarly, $\hat{\sigma}_{|f'(t, \hat{\theta}_l^*)|}$, $l = 1, \dots, B_1$, is calculated for each of the generated first-level bootstrap samples using the second-level bootstrap samples. With these values

$$D^{*,l} := \max_{t \in T} \frac{|f'(t, \hat{\theta}_l^*)| - |f'(t, \hat{\theta})|}{\hat{\sigma}_{|f'(t, \hat{\theta}_l^*)|}}$$

can be calculated for each $l = 1, \dots, B_1$. The respective empirical $(1 - \alpha)$ -quantile of the distribution yields an estimate of the critical value c . A visual representation of this process is provided in the blue rectangle in Figure J (see supplementary material).

3 Simulation study

3.1 Setting and data generation

In order to evaluate the performance of our newly proposed method, a simulation study was conducted. Motivated by Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022) and Kappenberg, F. and Grinberg, M. and Jiang, X. and Kopp-Schneider, A. and Hengstler, J.G. and Rahnenführer, J. (2021) we focused on the 4pLL and the beta model as representatives of the models discussed in Section 4. This choice was also based on the findings regarding the 4pLL/sigmoid Emax model in Duda, J.C. and Kappenberg, F. and Rahnenführer, J. (2022). The beta model was chosen as a suitable non-monotonous alternative. The simulation settings were inspired by the Western diet mice trial discussed in Section 4. Precisely, we used the parameters obtained from the expression of the genes 'Cd163' (4pLL model, see Section 4.1 for details) and 'Fam83a' (beta model) in WD-fed mice.

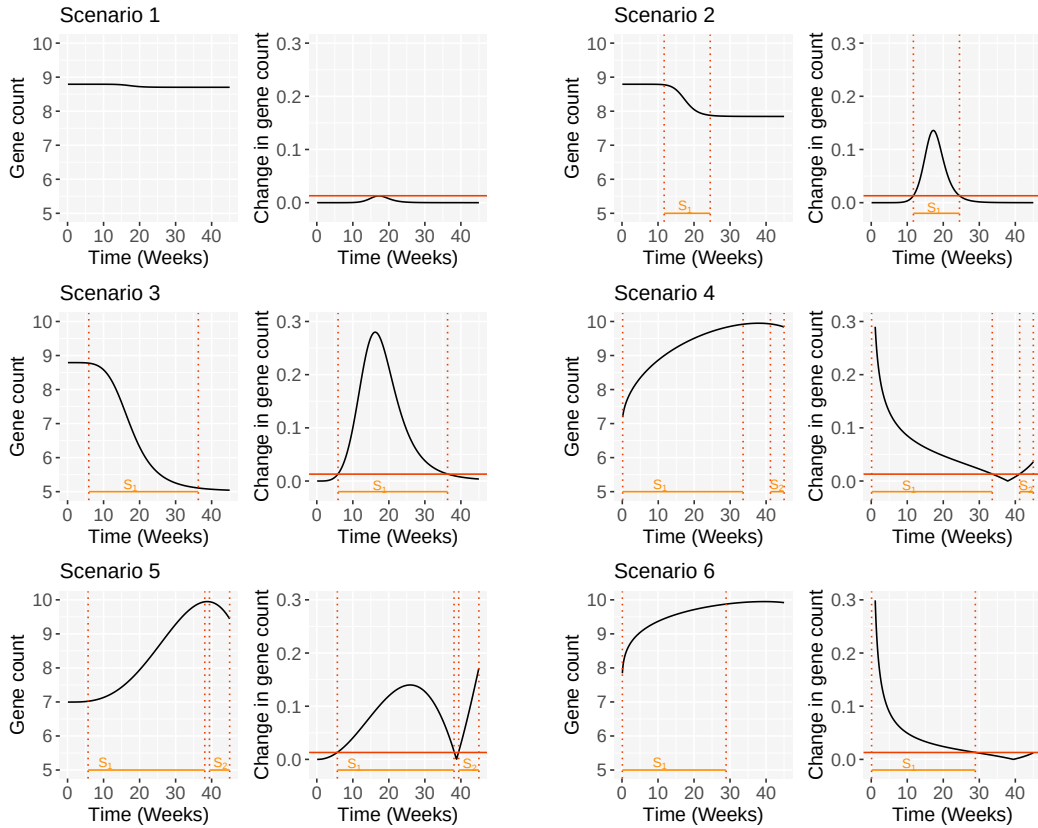


Figure 2: Visualization of the different scenarios. The horizontal red line indicates the relevance threshold $\lambda = \log_2(1.5) \times 45^{-1}$. Each case depicts the true model on the left and the true first derivative on the right. The vertical red lines mark the time frame of relevant change in regards to λ .

This resulted in the six basic scenarios, which are summarized in Table 1 and depicted in Figure 2. Scenarios 1-3 were generated using the 4pLL model, while the latter three were generated with the beta model. In Scenario 1, there is no relevant change in gene expression respective to the threshold $\lambda = \log_2(1.5) \times 45^{-1} \approx 0.0130$. Therefore, H_0 in (1) cannot be rejected. In all other scenarios, there is a change in expression larger than λ , and, therefore, these scenarios correspond to the situation under the alternative. Scenario 4 reflects the

unaltered parameters as estimated. Scenarios 2, 3, 5 and 6 were based on the real examples, but changed to reflect a different slope, see Table 1.

Table 1: Overview of simulation scenarios.

Sce.	Short name	Model	Parameters θ	Description
1	No relevant change	4pLL	$a = 8.791, b = -0.089$ $c = 17.589, h = 10.000$	Never exceeds $\lambda = 0.0130$ Reflects the margin of H_0 .
2	Small Jump	4pLL	$a = 8.791, b = -0.946$ $c = 17.589, h = 10.000$	Exceeds $\lambda = 0.0130$ $\tilde{S} = (11.7, 24.5)$
3	Large Jump	4pLL	$a = 8.791, b = -3.783$ $c = 17.589, h = 5.000$	Exceeds $\lambda = 0.0130$ $\tilde{S} = (5.9, 36.3)$
4	Dip	beta	$a = 6.997, b = 2.952$ $\delta_1 = 0.506, \delta_2 = 0.215$	Exceeds $\lambda = 0.0130$ twice $S_1 = (0, 33.6), S_2 = (41.2, 45)$
5	Dip alternative	beta	$a = 6.997, b = 2.952$ $\delta_1 = 3.286, \delta_2 = 1.290$	Exceeds $\lambda = 0.0130$ twice $S_1 = (5.7, 38.1), S_2 = (39.4, 45)$
6	No Dip	beta	$a = 6.997, b = 2.952$ $\delta_1 = 0.228, \delta_2 = 0.084$	Exceeds $\lambda = 0.0130$ $\tilde{S} = (0, 28.9)$

The observations were simulated at seven time points (0, 3, 9, 15, 21, 27, 33, 39 and 45 weeks) in accordance with the feeding times observed in the Western diet mice trial, following the beginning of the study. A total of 47 data points were simulated consisting of 5 mice at the first 5 time points and 4 and 8 mice at 39 and 45 weeks, respectively.

Of note, the data set used for the simulation was derived from a toxicological study that required the sacrifice of the mice in order to obtain measurements. In order to evaluate the disease progression, sacrifices were made at different time points. Consequently, each measurement represents a distinct mouse and we can assume independence across the measurements. Such studies are referred to as *serial-sacrifice experiments* (Dewanji, A., 2005; Bergman, S.W. and Turnbull, B.W., 1983) and will be the focal point of the ensuing sections. As previously stated in Section 2.1, the procedure can then be simplified by assuming the error terms to be independent normally distributed $\varepsilon_{p,q} \sim N(0, \sigma^2)$. Given the relatively small number of observations ($n = 47$) we will apply the unbiased estimator for the variance given by

$$\hat{\sigma}^2 = \frac{1}{n-r} \sum_{p=1}^m \sum_{q=1}^{n_p} \{y_{p,q} - f(t_p, \hat{\theta})\}^2.$$

In order to gain a comprehensive understanding of the impact of the standard deviation, five distinct levels for each of the six basic scenarios were considered: 'small', 'mid-small', 'medium', 'mid-large' and 'large'. The 'medium' standard deviation of the realistic Scenario 4 is as estimated from the real world data. For all other scenarios, the 'medium' standard deviation was calculated via linear transformation with regard to the respective maximum absolute differences in gene expression compared to the realistic example of the same model type. Regarding the parameterization introduced in Section 2.1, the maximum absolute value corresponds to the parameter $b \in \theta$. Accordingly, we designate $b_{\text{sce}} \in \theta_{\text{sce}}$ as the maximum absolute difference of the scenario and $b_{\text{real}} \in \theta_{\text{real}}$ as the maximum absolute difference estimated from the example. With this, we obtained the transformed standard deviation by $\sigma_{\text{sce,medium}} = b_{\text{sce}} \times b_{\text{real}}^{-1} \times \sigma_{\text{real}}$, where σ_{real} the estimated value from the

real world data.

From these 'medium' values of standard deviation, the 'small', 'mid-small', 'mid-large' and 'large' levels were calculated by multiplying with 0.5, 0.75, 1.5 and 2, respectively, see Table A in the supplementary material for details.

Given either the model function, f , of the beta model or the 4pLL model and the values of θ and σ , the gene expression data was simulated by drawing iid. errors $\varepsilon_{p,q}^{\text{sim}} \sim N(0, 1)$ and applying the formula $y_{p,q}^{\text{sim}} = f(t_p, \hat{\theta}) + \hat{\sigma} \varepsilon_{p,q}^{\text{sim}}$ for $p = 1, \dots, m$ and $q = 1, \dots, n_p$. As described in Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022) and following the postulations from Efron, B. and Tibshirani, R.J. (1994) we chose a large value for the estimation of the quantiles by the first-level bootstrap and a relatively small value for the estimation of the standard error by the second-level bootstrap. Precisely, we chose $B_1 = 500$ and $B_2 = 25$ repetitions. Smaller values would lead to imprecise estimations, whereas larger values could improve the results, but would require excessive computational effort. Each procedure was repeated in 1,000 simulation runs. Outcomes of the simulation study were the number of rejections of H_0 in (1) and the identified time frame of significant change in gene expression. The majority of the code was run on R version 4.3.2 using a Windows 11 x64 PC. The simulation scenarios were run on a HPC-cluster employing R version 4.3.1. The runtime of each simulated study on a commercially available Windows PC is approximately 5 minutes, with a variability of a few minutes depending on the number of CPU cores.

3.2 Results

3.2.1 Rejection of H_0

Table 2 shows the number of rejections of H_0 in 1,000 runs for all configurations. The left column in Figure 3 depicts the estimated confidence bands for a mid-small and mid-large standard deviation. We refer to the supplementary material Figures M-O for the other cases.

Table 2: Number of rejections of H_0 in 1,000 runs.

σ level	Sce. 1	Sce. 2	Sce. 3	Sce. 4	Sce. 5	Sce. 6
small	0	999	1000	1000	1000	1000
m.-small	0	946	818	1000	1000	1000
medium	0	830	514	1000	1000	1000
m.-large	0	627	401	989	999	978
large	0	372	303	814	957	780

In Scenario 1, no false rejection of H_0 was observed in any of the cases, indicating a type I error of 0. This scenario corresponds to the margin of the null hypothesis in (1), here given by $\lambda = 0.0130$. The conservatism of the test decision even for a large standard deviation may therefore be partially attributed to the effect being very small and an overall tendency to slightly underestimate the time frame of significant change in expression discussed in the next subsection.

In general, our proposed method shows high power when the beta model is the true model. Up to a medium standard deviation, H_0 is always rejected for Scenarios 4, 5 and 6. Even for a large standard deviation, the power remains at about 80% for Scenarios 4 and 6 and above 95% for Scenario 5. Accordingly, the estimated confidence bands show a low variability.

For the remaining scenarios generated with the 4pLL model – Scenario 2 and 3 – we observe a large influence of the standard deviation. In particular, the power is above 80% for a small and mid-small variability, but lies between 30% and 65% for a large and mid-large variability, respectively (see 2). This influence seems to be more

pronounced for the larger jump in gene expression simulated in Scenario 3. In this case, only 514 of significant changes were detected for a medium standard deviation.

It is noteworthy that the (mid-)large standard deviations (see Table A in the supplementary material) employed in the generation of Scenarios 2 and 3 is considerably large compared to the absolute change in gene expression. Naturally, this leads to a high variability in the generated simulation studies (see Figure K in the supplementary material). However, we observe that in the case of the 4pLL model, the translation of a high degree of variability on the count scale leads to an extreme degree of variability on the scale of the change in gene expression considered when applying the first derivative (see Figure L in the supplementary material). It is important to interpret the results presented in light of these considerations. It is also important to note that the same effect occurs when generating the first and second level bootstrap runs. This is likely to have an impact on the estimation process, especially if the data set has a high degree of variability.

For Scenarios 2 and 3 we note a visible crater in the estimated confidence bands around the week where the true first derivative reaches its maximum, which corresponds to the inflection point of the model function (see Figure 3 and Figures M-O in the supplementary material). While the extent is likely to be at least partially attributable to the size of the standard deviation, the trend is discernible across all cases. Given that the effect appears to be more pronounced in Scenario 3 than in Scenario 2, it is plausible that its magnitude may correlate with the steepness of the true model function in the area around its inflection point. This conclusion is supported by the absence of a comparable crater for Scenario 5, the only scenario with an inflection point generated using the beta model, which is markedly less steep than Scenarios 2 and 3.

3.2.2 Estimated time frame of significant change in gene expression

While the number of rejections of H_0 is largely satisfactory, a reliable detection of the time frame of significant change in gene expression is of greater interest. Therefore, we considered the bias and variance of each start/end time point of the estimated time frame of significant change in gene expression. Furthermore, we examined the number of runs in which the true number of coherent subsets of $\tilde{S}$ was correctly identified. We refer to Table B in the supplementary material for a summary of all results. As anticipated, the true number of coherent subsets was consistently detected for Scenarios 2 and 3 if H_0 was rejected. Given the monotonicity of the 4pLL model, the time frame of significant change in expression is, by definition, continuous. However, as described above, a crater was observed in some estimated lower simultaneous confidence bands. In a small number of cases, this resulted in the identification of two disjointed subsets. Overall, the number of these cases is less than 12 for Scenario 2. In case of Scenario 3, the same is observed except for a mid-small and medium standard deviation, which respectively fall within the ranges of 200 and 70 cases.

While H_0 was reliably rejected for Scenarios 4 and 5, the second time frame of significant change in expression was not always detected. The smaller dip in Scenario 4 was only identified approximately 120 times when the standard deviation was small and approximately 40 times otherwise. The more pronounced dip in Scenario 5 was detected 999 times for a small standard deviation, 671 times for a medium standard deviation and 260 times for a large standard deviation. However, it should be noted that in both scenarios, the maximum value in gene count is reached around week 39, which is the second to last sample time point. Therefore, the detection of the dip is contingent only on the last measurement. In Scenario 6, a dip at the end was incorrectly identified less than 15 times in all cases.

The right columns in Figures 3 depict the estimated time frames of significant change in gene expression for each of the 1,000 simulation runs for the mid-small and mid-large standard deviations, respectively. We refer to Figures M-O in the supplementary material for the other cases. The true values are indicated by the red lines. In general, we note that the estimated start of the time frame of significant change in expression displays a positive bias for all scenarios, whereas the estimated end displays a negative bias. Therefore, the time frame is overall slightly underestimated. The bias depends largely on the standard deviation and the slope of the first

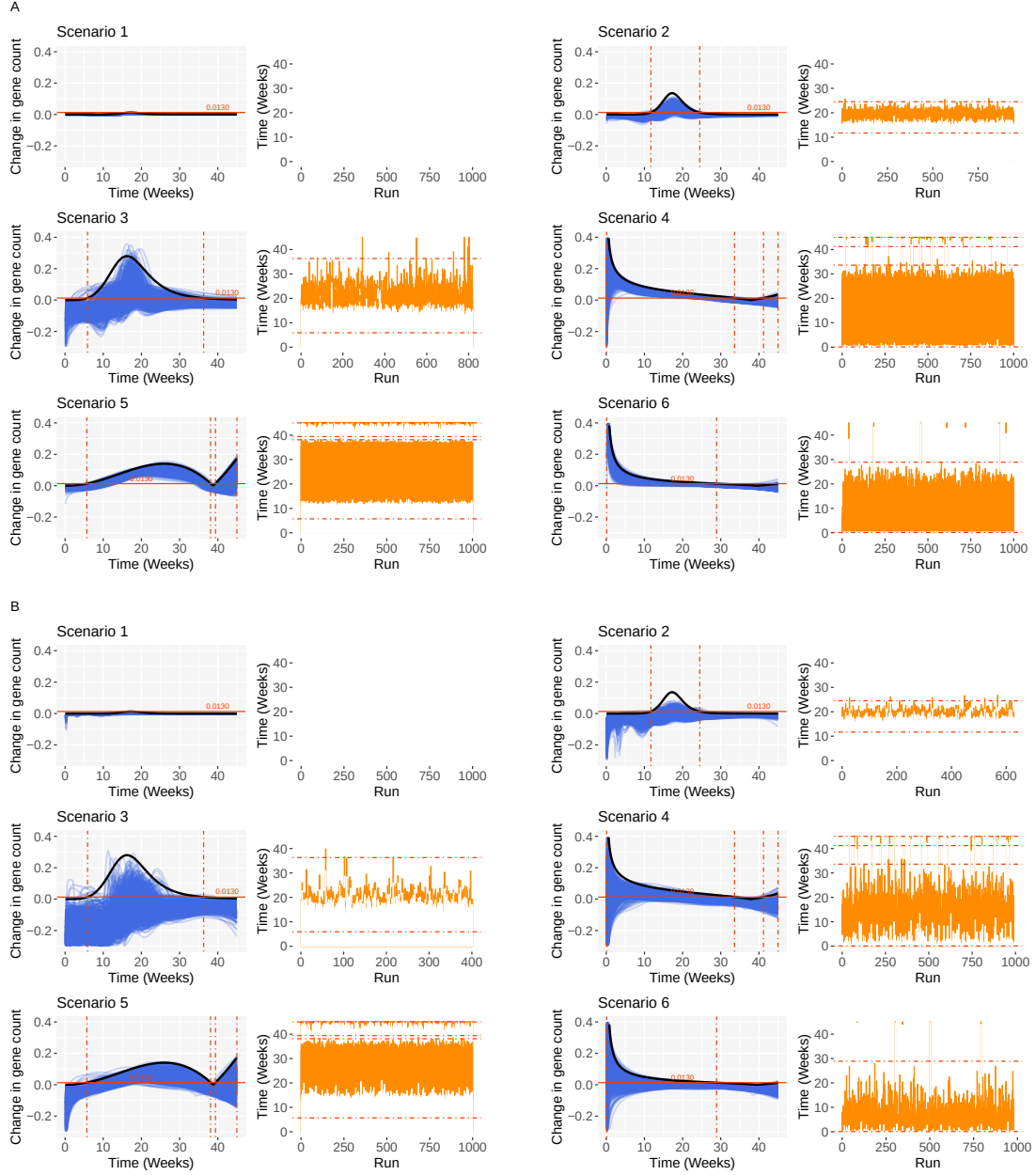


Figure 3: Results of the six simulation scenarios for a mid-small (A) and mid-large (B) standard deviation. The left column depicts the true first derivative and the estimated confidence bands. The right column depicts the estimated time frames of significant change in gene expression for each run.

derivative. Scenario 6 exhibits the smallest bias for the onset of the significant change at the beginning of the study, never exceeding 0.375 weeks. Scenario 4 is similarly flat. Its bias never exceeds 2.379. Scenario 5 is characterised by a steeper slope and exhibits a noticeably larger bias. The range is from 3.956 weeks for a small standard deviation to 8.865 weeks for a large standard deviation. For Scenarios 2 and 3 – disregarding the crate when detected – we observe similar results. The influence of the slope seems to have a reverse effect on the bias of the end of the identified time frame of significant change. In particular, flat scenarios, such as Scenario 6, demonstrate a noticeably larger bias for all levels of standard deviation. In addition, the bias is overall larger than for the start of the significant change, with a range of -1.225 to -23.591 weeks.

In Scenarios 4 and 5, a second time frame of significant change in expression was simulated. Since in both cases it is cut off at the end of the study, we consider only the starts of the second subsets of $\tilde{S}$, respectively. Again, we note a positive bias, which ranges from almost 2 weeks to nearly 3.5 weeks. Finally, the variance shows a similar behavior to the bias.

4 Case study

In this section, three possible applications of our method are presented. For illustration, we use data from the Western diet mice study (Ghallab, A. et al., 2021). A total of 79 male mice were fed a Western (WD) or standard (SD) diet for a maximum of 48 weeks to assess the effect of a high fat diet on the development of non-alcoholic fatty liver disease. Up to 8 mice per diet were sacrificed at 3, 6, 30, 36, 42 and 48 weeks, and WD-fed mice were additionally sacrificed at 12, 18 and 24 weeks. RNA was extracted from frozen liver tissue and RNA-seq analysis was performed. Here, we use the pre-processed data set, that was generated using Salmon (Patro, R. and Duggal, G. and Love, M. I. and Irizarry, R. A. and Kingsford, C., 2017) and the R packages tximeta (Love, M.I. and Soneson, C. and Hickey, P.F. and Johnson, L.K. and Pierce, N.T. and Shepherd, L and Morgan, M. and Patro, R, 2020) and DESeq2 (Love, M.I. and Huber, W. and Anders, S., 2014). The count data was normalized and $\log_2$ -transformed using the *vst()* function (Tibshirani, R., 1988; Huber, W. and von Heydebreck, A. and Suelmann, H. and Poustka, A. and Vingron, M., 2003; Anders, S. and Huber, W., 2010)) from the DESeq2 package. As the initial samples were collected at 3 weeks, the beginning of the study at $t = 0$ in the subsequent analysis represents these initial data points.

The first example illustrates the application of the method to four genes expressing different shapes. For each gene, the methodology is demonstrated using a parametric model fit as well as fitting B-splines. In the second example, we select a pool of nearly 10000 genes and use our method to identify those that change significantly between 10 and 25 weeks. A GO enrichment analysis is then performed on the identified genes. In all cases, the two-step bootstrap procedure was implemented with $B_1 = 500$ and $B_2 = 25$ bootstrap runs as described in Section 3.1.

4.1 Application to an individual gene

For this example we applied the proposed method to four individual genes measured in WD-fed mice. The method was implemented using two different types of fit: a parametric model fit and a B-spline fit to the data. Firstly, the results obtained from the application of the parametric models will be presented. In a preceding model selection step, the 4pLL model, the beta model, the exponential model and the quadratic model were fitted to the gene expression data, respectively. The chosen examples are such that each of the parametric models is represented by one of the genes. Precisely, the 4pLL model is illustrated by the gene 'Cd163' (upper left), the beta model is illustrated by the gene 'Fam83a' (upper right), the exponential model is illustrated by the gene 'Dbp' (lower left) and the quadratic model is illustrated by the gene 'Tm7sf2'. The left panels of each gene in Figure 4 depict the respective estimated gene expression curves.

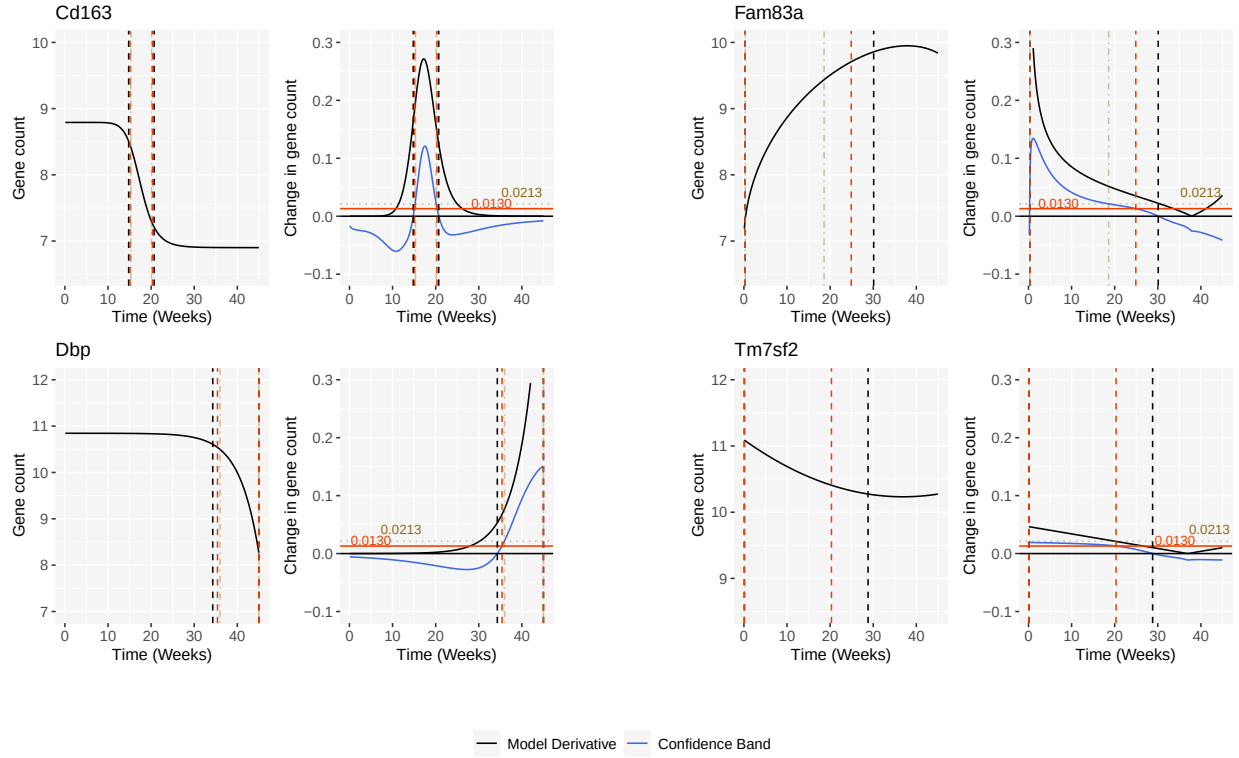


Figure 4: On the left is a depiction of the fit of the best parametric model to the normalized gene count over time for each of the four example genes. On the right, the first derivative and the estimated confidence band of the model fit are displayed. The horizontal red line indicates $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines indicate the corresponding time frame of significant change in gene expression. The yellow dotted lines correspond to an alternative choice of $\lambda = \log_2(1.5) \times 27.5^{-1}$. The black lines represent the case $\lambda = 0$. The expression of the gene 'Cd163' follows a 4pLL model (upper left), the expression of the gene 'Fam83a' follows a beta model (upper right), the expression of the gene 'Dbp' follows an exponential model (lower left) and the expression of the gene 'Tm7sf2' follows a quadratic model (lower right).

In Ghallab, A. et al. (2021) both 'Cd163' and 'Fam83a' were identified as RJG. This implies that the gene in question is only deregulated after a specific period of time, in these cases, after 24 and 18 weeks, respectively. According to identification as an RJG, it appears that approximately 10 – 12 weeks pass before the gene 'Cd163' is down-regulated. Following an additional 10 – 12 weeks, the deregulated expression seems to reach a constant value. In contrast, the gene 'Fam83a' appears to become deregulated immediately. After week 20 – 30, the deregulated expression seems to reach a constant value. Although not identified as RJG in Ghallab, A. et al. (2021), the 'Dbp' gene appears to become down-regulated only after a period of 30 – 35 weeks. In contrast, the 'Tm7sf2' gene appears to be deregulated at the beginning of the study, reaching a plateau after week 30.

In the respective right panels of Figure 4, the corresponding first derivatives and the estimated confidence bands are depicted. The various horizontal lines represent different potential choices for the threshold λ , while the corresponding vertical lines represent the respective time frames of significant change in gene expression. The black line represents the case where $\lambda = 0$, while the red line corresponds to the aforementioned case $\lambda = \log_2(1.5) \times 45^{-1} = 0.0130$ (see Section 2.2). Similarly, a third threshold was calculated, indicated by the yellow line, which corresponds to a change of $\log_2(1.5)$ in gene expression over the span of half the study. This third threshold is given by $\lambda = \log_2(1.5) \times 27.5^{-1} = 0.0213$.

We observe that H_0 is rejected for all genes and all thresholds, with the exception of the gene 'Tm7sf2'. For this gene, a significant change cannot be concluded at the threshold of $\lambda = 0.0213$. In the following, we will present a summary of the analysis for $\lambda = \log_2(1.5) \times 45^{-1}$ as an illustrative example. For a detailed analysis of all four genes we refer to the supplementary material. Confidence intervals for the time points of interest were estimated using a third-level bootstrap (for details on the exact procedure we refer to the supplementary material).

The period during which the estimated significant changes in gene expression of 'Cd163' occurs begins at 15.1 weeks (95% CI [13.2, 17.8]) and ends at 20.3 weeks (95% CI [17.4, 23.5]). The 'Dbp' gene shows similar delayed deregulation. The estimated time frame of significant changes in gene expression begins at 35.4 weeks (95% CI [29.3, 37.0]) and continues until the end of the study at 45 weeks (95% CI [41.4, 45.0]). Consequently, we can conclude the classical RJG form for both genes. For the gene 'Fam83a', the estimated time period begins at 0.3 weeks (95% CI [0.1, 1.9]) and ends at 24.9 weeks (95% CI [15.7, 32.1]). Similarly, the time frame of significant change in expression estimated for the gene 'Tm7sf2' begins at the start of the study (95% CI [0.1, 14.0]) and ends at 20.3 weeks (95% CI [12.2, 25.6]). Therefore, we would not conclude an RJG form in both cases.

Additionally, we applied the proposed method using B-splines to all four genes and examined the influence of both the degree of the basis functions and the number of knots. Precisely, we conducted the analysis using the R packages *splines2* (Wang, W. and Yan, J., 2025, 2021), which includes the derivatives of the basis functions. The degree of the basis functions was selected to be $l \in \{2, 3\}$ with either two or three internal knots, which were automatically set at suitable quantiles (Wang, W. and Yan, J., 2025). Basis functions of higher degree and a greater number of knots were considered, but overall resulted in overfitting for the given data. We refer to Figure A in the supplementary material for a depiction of the basis functions and their first derivatives. After calculating the values of the B-spline basis using the R function *bspline()*, a linear model was fitted using the *lm()* function. The first derivatives were calculated via the function *dfs()*.

For the gene 'Cd163', the results obtained are mostly comparable to those produced by fitting parametric models. The time frame of significant change overlaps for all cases. In case of $l = 2$ degrees with two knots, the results for the gene 'Fam83a' are also very similar to the parametric case. However, for all other B-spline cases, we detect a period of inactivity around 10 weeks. This phenomenon can be attributed to the flexible nature of B-splines and the noticeable dip in expression presented by the measures at week 9. However, the estimated time frames for significant changes in gene expression are found to overlap satisfactorily for these three cases, reinforcing the robustness of the results obtained. For the gene 'Dbp' two distinct time frames of significant change in expression were detected. While the latter period is comparable to the estimated period applying the

Cd163

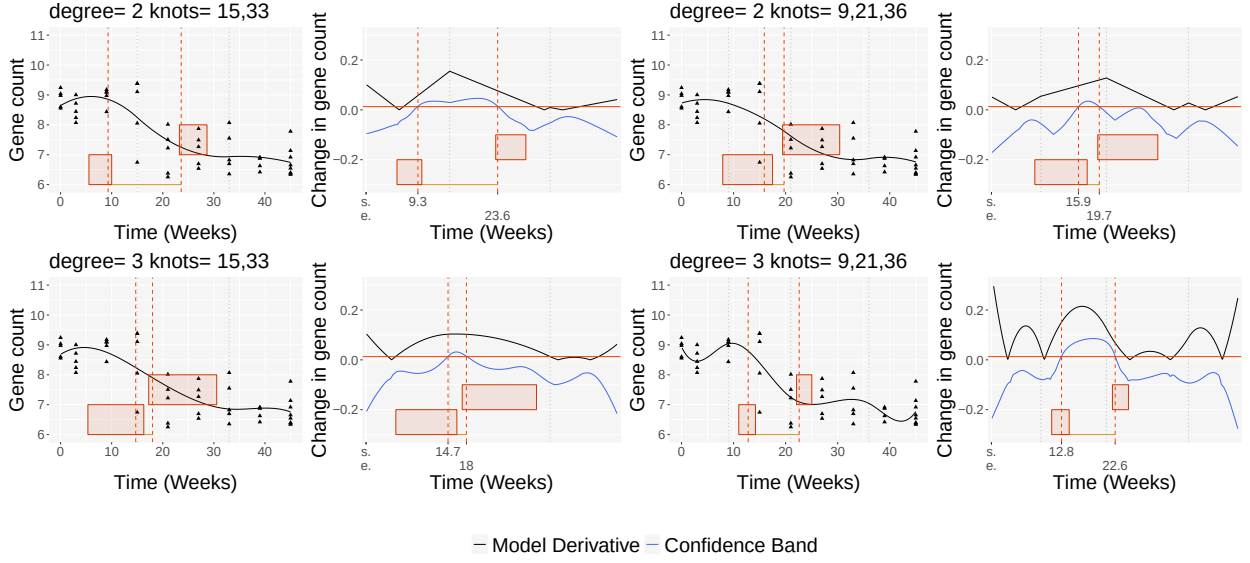


Figure 5: The fit of B-splines to the normalized gene count of the gene 'Cd163' over time is depicted on the left for each presented combination of degree and number of knots. The degree of the basis functions and the knot placement are stated in the subtitle of the respective plots. Additionally, the knots are indicated by gray dotted lines. On the right, the first derivatives and the estimated confidence bands are displayed. The horizontal red lines indicate $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines and the horizontal yellow lines indicate the corresponding time frames of significant change in gene expression. The red boxes illustrate the estimated confidence intervals for the respective time points of interest.

parametric approach, the earlier period is not detected, possibly indicating an insufficient fit of the exponential model and thereby the collection of parametric models. Interestingly, we detect no significant change for the expression of the gene 'Tm7sf2'. However, the detected signal found utilizing a parametric fit is very small (see Figure G in the supplementary material). This suggests that the parametric approach may be slightly more sensitive when the parametric model fits the data well.

The findings indicate that with increasing degree and number of knots, the estimated confidence bands become more wavy in nature, as is also observed for the model fit applying B-splines. This noticeable structure is clearly influenced by the placement of the B-spline basis functions along the knots and is expected due to the definition. We note that for the three genes 'Cd163', 'Fam83a' and 'Dbp' the estimated time frames are the widest for $l = 2$ degrees with two knots. However, it is not possible to make a similar assertion regarding the remaining cases. In general, the correspondence of the fits indicates that core regions of activity can be detected by the proposed method, irrespective of the specified preliminaries.

Figure 5 depicts the different fits of B-splines for the gene 'Cd163' on the left and their respective first derivatives and estimated confidence bands on the right. As before, the red horizontal lines indicate the threshold $\lambda = 0.0130$. The red vertical lines indicate the estimated periods of change in gene expression. In general, the fits with B-splines appear less steep than the parametric fit with the 4pLL model. In accordance with this observation, a significant change in gene expression was detected from 9.3 – 16.2 weeks until 17.2 – 22.4 weeks. This time frame is slightly broader for degree $l = 2$ with two knots or $l = 3$ with three knots than the outcomes derived from employing the parametric approach.

We refer to the supplementary material for an analysis of the results obtained by fitting B-splines to the three other genes. Moreover, an application for a gene that deregulates in both diet groups and a subsequent comparison of the estimated time frames of significant changes in gene expression is included in the supplementary

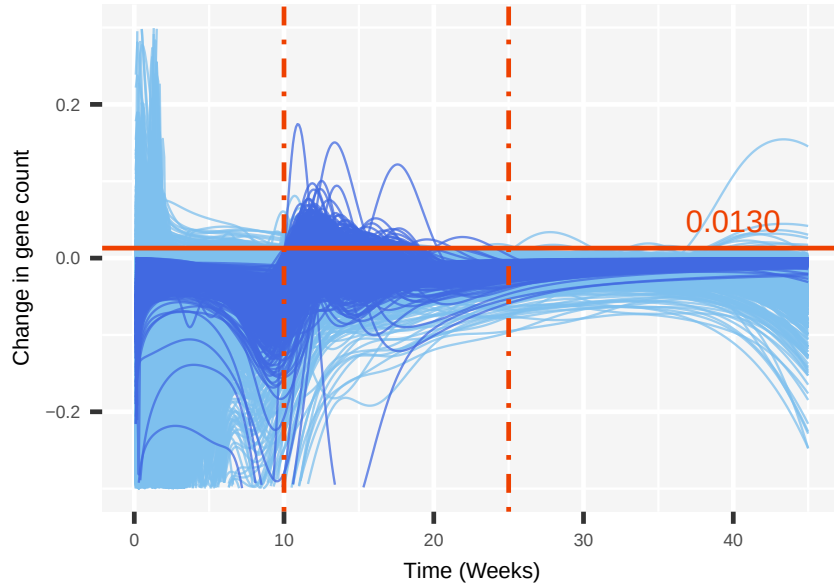


Figure 6: Estimated confidence bands of the 9881 pre-selected genes. The darker blue lines correspond to the genes for which the estimated period of significant change in expression was determined to be between 10 and 25 weeks.

material.

4.2 GO enrichment analysis

For this example we applied our proposed method to identify a subset of differentially expressed genes that deregulate during a pre-selected time period (10 to 25 weeks). In a subsequent step, we conducted a GO enrichment analysis on these genes. This means, we examined if a significant number of the identified genes were assigned to pre-specified GO groups (Consortium, G. O., 2001), in which the genes were grouped according to their connection to biological functions. This approach would permit a practitioner to draw inferences regarding the time frame within which specific biological processes occur. For instance, this could be used to evaluate the progression of a disease.

To achieve this, we selected 9881 genes that were identified as differentially expressed in Ghallab, A. et al. (2021) comparing SD-fed mice at week 0 (time on study) and WD-fed mice at week 45.

We assumed that the 4pLL model was the most appropriate for the analysis and applied our method to identify the time periods of significant change in gene expression. Figure 6 depicts the estimated confidence bands. The results were then employed to identify the genes with an estimated time period of significant change in expression between weeks 10 and 25. A total of 569 genes were identified.

A GO enrichment analysis was conducted on the aforementioned genes using the topGO R package (Alexa, A. and Rahnenführer, J. , 2024; Alexa, A. and Rahnenführer, J. and Lengauer, T., 2006). In essence, Fisher’s exact test is employed to verify whether a number of interesting genes is annotated to a specific functional group that exceeds what would be expected by chance. The 20 gene ontology groups with the smallest adjusted p-value are presented in Table C in the supplementary material. For instance, the majority of the genes (525) were assigned to the GO group ‘inflammatory response’. This is in accordance with the findings that inflammation is a pivotal event in the progression of NAFLD (Ghallab, A. et al., 2021; Sutti, S. and Albano, E. , 2020). In Ghallab, A. et al. (2021) few inflammatory foci were observed in week 3 and the number strongly increased

after week 36. Lipogranulomas, single or multiple fat globules surrounded by chronic inflammatory cells and Kupffer cells (Brunt, E.M., 2002), were observed as early as week 6 and increased at week 18.

5 Discussion and conclusion

In this paper, we proposed a parametric method to estimate the time frame during which the expression of a gene changes significantly. Thereby, we have provided a tool, that allows for an interpretation that is not limited to the onset of an effect or the observed time points and produces results based on a hypothesis test to the α -level. As demonstrated by the 4pLL model, the beta model, the exponential model, the quadratic model and B-splines, the approach is applicable to a wide range of potential models, with the singular restriction that the first derivative can be derived analytically from the model function. It can be employed to analyse the expression of a single gene or, in conjunction with a GO enrichment analysis, to infer biological function. Moreover, the methodology can be applied to other forms of time-response or dose-response data.

Of course, there are limits to the method. The simulation results discussed in Section 3.2 show a clear tendency to underestimate the time frame of significant change in expression. The effect is relatively modest for medium levels of data variability. However, it becomes more pronounced with greater variability. Overall, the method would benefit from a built-in stabilization procedure against the predictable effects of a large standard deviation. Similarly, the noticeable effect of the form of the B-spline basis functions on the estimated confidence band may complicate the interpretation of the results in cases of overfitting. One way to combat this is to include penalization in the method, such as by fitting P-splines instead (see (Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D., 2022)). In addition, it would be interesting to explore the possibility of adapting alternative construction methods for simultaneous confidence bands to the first derivative. This is especially true for B-splines, since the first derivative of the entire polynomial spline is again a linear combination of B-spline basis functions. Therefore, established methods for the estimation of simultaneous confidence bands could be applied as an alternative, e.g. Krivobokova, T. and Kneib, T. and Claeskens, G. (2010) for penalized splines. Also, while we attempted to discern any relationship between the estimated time frames and knot placement in the case study, a thorough consideration of this topic would exceed the scope of this paper. However, it would be an interesting and important endeavour to expand this analysis in the future.

Often, a statistical conclusion on more than one gene expression curve is of interest. In Section 4.2, we presented a possible application to a large number of genes in combination with a GO enrichment analysis. Nevertheless, a multitude of potential biological questions may necessitate alternative approaches. This frequently gives rise to a multiple testing problem. Hence, our proposed method would highly benefit from the development of a method, that corrects for this issue beyond a simple correction of the p -values.

In conclusion, we posit that our proposed method is a valuable addition to the existing tools for the analysis of gene expression data. It allows for an comprehensive analysis over the entire study period and is applicable to a wide range of expression curves.

Supplementary material

The case study data set is publicly available at the SRA database with reference number PRJNA953810. Supplementary material providing additional tables and figures can be found online. Corresponding R code, which can be used to reproduce the analysis of the case study and the simulations, is available at <https://github.com/LuciaAmeis/Identification-of-changes-in-gene-expression>.

Funding

This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116).

References

- Alexa, A. and Rahnenführer, J. (2024). topGO: Enrichment Analysis for Gene Ontology. R package version 2.56.0.
- Alexa, A. and Rahnenführer, J. and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600 – 1607.
- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*.
- Androulakis, I.P. and Yang, E. and Almon, R.R. (2007). Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, 9(1):205–228.
- Bergman, S.W. and Turnbull, B.W. (1983). Efficient sequential designs for destructive life testing with application to animal serial sacrifice experiments. *Biometrika*, 70(2):305–314.
- Bornkamp, B. and Pinheiro, J. and Bretz, F. (2009). MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies. *Journal of Statistical Software*, 29(7):1 – 23.
- Brunt, E.M. (2002). Alcoholic and nonalcoholic steatohepatitis. *Clinics in Liver Disease*, 6(2):399 – 420.
- Consortium, G. O. (2001). Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8).
- Delignette-Muller, M.-L. and Forfait, C. and Billoir, E. and Charles, S. (2011). A new perspective on the Dunnett procedure: Filling the gap between NOEC/LOEC and ECx concepts. *Environmental Toxicology and Chemistry*, 30(12):2888–2891.
- Dewanji, A. (2005). *Serial-Sacrifice Experiments*. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd.
- Duda, J.C. and Kappenberg, F. and Rahnenführer, J. (2022). Model selection characteristics when using mcp-mod for dose-response gene expression data. *Biometrical Journal*, 64(5):883–897.
- Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. CRC Press, 1st ed. edition.
- Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D. (2022). *Regression: Models, Methods and Applications*. Springer Berlin, Heidelberg, 2nd edition.
- Ghallab, A., Myllys, M., Friebe, A., Duda, J., Edlund, K., Halilbasic, E., Vucur, M., Hobloss, Z., Brackhagen, L., Begher-Tibbe, B., Hassan, R., Burke, M., Genc, E., Frohwein, L.J., Hofmann, U., Holland, C.H., González, D., Keller, M., Seddek, A., Abbas, T., Mohammed, E.S.I., Teufel, A., Itzel, T., Metzler, S., Marchan, R., Cadenas, C., Watzl, C., Nitsche, M.A., Kappenberg, F., Luedde, T., Longerich, T., Rahnenführer, J., Hoehme, S., Trauner, M., and Hengstler, J.G. (2021). Spatio-Temporal Multiscale Analysis of Western Diet-Fed Mice Reveals a Translationally Relevant Sequence of Events during NAFLD Progression. *Cells*, 10(10).
- Hothorn, L.A. (2014). Statistical evaluation of toxicological bioassays - a review. *Toxicology Research*, 3(6):418–432.

- Huber, W. and von Heydebreck, A. and Suelmann, H. and Poustka, A. and Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical applications in genetics and molecular biology*.
- Jensen, S.M. and Kluxen, F.M. and Ritz, C. (2019). A review of recent advances in benchmark dose methodology. *Risk Analysis*, 39(10):2295–2315.
- Kappenberg, F. and Duda, J.C. and Schürmeyer, L. and Gül, O. and Brecklinghaus, T. and Hengstler, J.G. and Schorning, K. and Rahnenführer, J. (2023). Guidance for statistical design and analysis of toxicological dose–response experiments, based on a comprehensive literature review. *Archives of Toxicology*, 97(10):2741–2761.
- Kappenberg, F. and Grinberg, M. and Jiang, X. and Kopp-Schneider, A. and Hengstler, J.G. and Rahnenführer, J. (2021). Comparison of observation-based and model-based identification of alert concentrations from concentration-expression data. *Bioinformatics*, 37(14):1990 – 1996.
- Khatoon, Z. and Figler, B. and Zhang, H. and Cheng, F. (2014). Introduction to RNA-Seq and its applications to drug discovery and development. *Drug development research*, 75(5):324–330.
- Krivobokova, T. and Kneib, T. and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, 105(490):852–863.
- Krug, A.K., Kolde, R., Gaspar, J.A., Rempel, E., Balmer, N.V., Meganathan, K., Vojnits, K., Baquié, M., Waldmann, T., Ensenat-Waser, R., Jagtap, S., Evans, R.M., Julien, S., Peterson, H., Zagoura, D., Kadereit, S., Gerhard, D., Sotiriadou, I., Heke, M., Natarajan, K., Henry, M., Winkler, J., Marchan, R., Stoppini, L., Bosgra, S., Westerhout, J., Verwei, M., Vilo, J., Kortenkamp, A., Hescheler, J., Hothorn, L., Bremer, S., van Thriel, C., Krause, K.-H., Hengstler, J.G., Rahnenführer, J., Leist, M., and Sachinidis, A. (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Archives of toxicology*, 87:123–143.
- Love, M.I. and Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550.
- Love, M.I. and Huber, W. and Anders, S. (2024). *Analyzing RNA-seq data with DESeq2*. Accessed July 17, 2024.
- Love, M.I. and Soneson, C. and Hickey, P.F. and Johnson, L.K. and Pierce, N.T. and Shepherd, L and Morgan, M. and Patro, R (2020). Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLOS Computational Biology*, 16:e1007664.
- Möllenhoff, K. and Schorning, K. and Kappenberg, F. (2022). Identifying Alert Concentrations Using a Model-Based Bootstrap Approach. *Biometrics*, 79(3):2076 – 2088.
- NCBI (1996). *PubMed*. Accessed July 07, 2024.
- Neumann, M.H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333.
- Patro, R. and Duggal, G. and Love, M. I. and Irizarry, R. A. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14:417–419.
- Stark, R. and Grzelak, M. and Hadfield, J. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656.

- Sutti, S. and Albano, E. (2020). Adaptive immunity: an emerging player in the progression of NAFLD. *Nature Reviews Gastroenterology & Hepatology*, 17:81 – 92.
- Tibshirani, R. (1988). Estimating Transformations for Regression via Additivity and Variance Stabilization. *Journal of the American Statistical Association*, 83(402):394 – 405.
- Wang, L. and Yang, L. (2010). Simultaneous confidence bands for time-series prediction function. *Journal of Nonparametric Statistics*, 22(8):999–1018.
- Wang, W. and Yan, J. (2021). Shape-restricted regression splines with R package splines2. *Journal of Data Science*, 19(3):498–517.
- Wang, W. and Yan, J. (2025). *splines2: Regression Spline Functions and Classes*. R package version 0.5.4.
- Wang, Z. and Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

Supplementary material to: Identification of changes in gene expression

L. Ameis¹ and K. Möllenhoff^{1*}

¹Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne

1 Parametric model functions and first derivatives

In the following section, we present the four parametric models used as examples in this paper, along with their respective first derivatives. The parameterizations and descriptions of the parameters are in accordance with the conventions noted in Bornkamp, B. and Pinheiro, J. and Bretz, F. (2009, 2017).

4pLL/sigEmax

E_0	Placebo effect
E_{max}	Asymptotic maximum change from placebo effect
ED_{50}	Dose giving half of the asymptotic maximum effect
h	Hill parameter, determining the steepness of the model at the ED50

$$f_{4pLL}(t, (E_0, E_{max}, ED_{50}, h)) = E_0 + E_{max} \frac{t^h}{ED_{50}^h + t^h}$$

$$f'_{4pLL}(t, (E_0, E_{max}, ED_{50}, h)) = \frac{E_{max} \times ED_{50}^h \times t^{h-1} \times h}{(ED_{50}^h + t^h)^2}$$

beta model

E_0	Placebo effect
E_{max}	Maximum effect
δ_1	Delta1 parameter
δ_2	Delta2 parameter

$$f_{beta}(t, (E_0, E_{max}, \delta_1, \delta_2)) = E_0 + E_{max} B(\delta_1, \delta_2) \left(\frac{t}{scal} \right)^{\delta_1} \left(1 - \frac{t}{scal} \right)^{\delta_2}$$

$$f'_{beta}(t, (E_0, E_{max}, \delta_1, \delta_2)) = \frac{E_{max} \times B_{\delta} \times t^{\delta_1-1} \times \left(1 - \frac{t}{scal} \right)^{\delta_2-1} \times (\delta_1 \times scal - (\delta_1 + \delta_2) \times t)}{scal^{\delta_1+1}}$$

Hereby, let $B_{\delta} = \{ (\delta_1 + \delta_2)^{\delta_1+\delta_2} \} \times \{ \delta_1^{\delta_1} \times \delta_2^{\delta_2} \}^{-1}$ and $scal$ denote a fixed scaling parameter.

Exponential model

E_0	Placebo effect
E_1	E1 parameter
δ	Delta parameter, controlling the convexity of the model

$$f_{exp}(t, (E_0, E_1, \delta)) = E_0 + E_1 \left(\exp \left(\frac{t}{\delta} \right) - 1 \right)$$

$$f'_{exp}(t, (E_0, E_1, \delta)) = E_1 \times \left(\exp \left(\frac{t}{\delta} \right) \times \left(\frac{1}{\delta} \right) \right)$$

Quadratic model

E_0	Placebo effect
β_1	Beta1 parameter
β_2	Beta2 parameter, controls, whether model is convex or concave

$$f_{quad}(t, (E_0, \beta_1, \beta_2)) = E_0 + \beta_1 t + \beta_2 t^2$$

$$f'_{quad}(t, (E_0, \beta_1, \beta_2)) = \beta_1 + 2\beta_2 t$$

2 B-spline basis functions and first derivatives

Let the interval $[t_1, t_m]$ denote the study period, and let $t_1 = \kappa_1 < \dots < \kappa_{\tilde{m}} = t_m$ be a suitable partition, e.g. equidistant or quantile based. We call κ_j , $j = 1, \dots, \tilde{m}$, the inner knots. As presented in Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D. (2022), the B-spline basis functions of degree l are recursively defined as follows. Let $d = l + \tilde{m} - 1$:

$$\begin{aligned} l = 0 \quad B_j^0(t) &= I(\kappa_j \leq t < \kappa_{j+1}) = \begin{cases} 1 & \kappa_j \leq t < \kappa_{j+1} \\ 0 & \text{otherwise,} \end{cases} \\ l > 1 \quad B_j^l(t) &= \frac{t - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(t) + \frac{\kappa_{j+1} - t}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(t) \text{ for } j = 1, \dots, d. \end{aligned}$$

Hereby, $I()$ denotes the indicator function. To ensure that this definition is well-defined for the entire study period, it is necessary to add $2l$ outer knots that extend outside the interval $[t_1, t_m]$. Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D. (2022) also provides the first derivative of a B-spline basis function of degree l :

$$\frac{\partial}{\partial t} B_j^l(t) = l \left(\frac{1}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(t) - \frac{1}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(t) \right)$$

Therefore, the following holds for f :

$$f(t) = \sum_j \gamma_j B_j^l(t) \implies f'(t) = l \sum_j \frac{\gamma_j - \gamma_{j-1}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(t).$$

The basis functions used for fitting B-splines in Section 4 of the main manuscript are depicted in Figure A.

3 On confidence intervals for the time points of interest

In Section 2.2 we obtained time points of significant change in gene expression. However, as estimators, they are subject to uncertainty. Thus, the construction of confidence intervals for these time points is of huge practical importance. The method proposed is an extension to the two-step bootstrap described in Section 2.3. Precisely, another bootstrap level is added in advance, and the two-step procedure described above is then applied to each sample. This provides an estimated lower confidence band for each of the generated data sets, allowing calculation of the time frames of significant change in gene expression and the time points of interest. In the following analysis, only those cases with the same number of coherent subsets as the original data set are considered. The corresponding time points of interest from all bootstrap samples are combined (e.g. all t_{start}^* and all t_{end}^* , where the * denotes the bootstrap

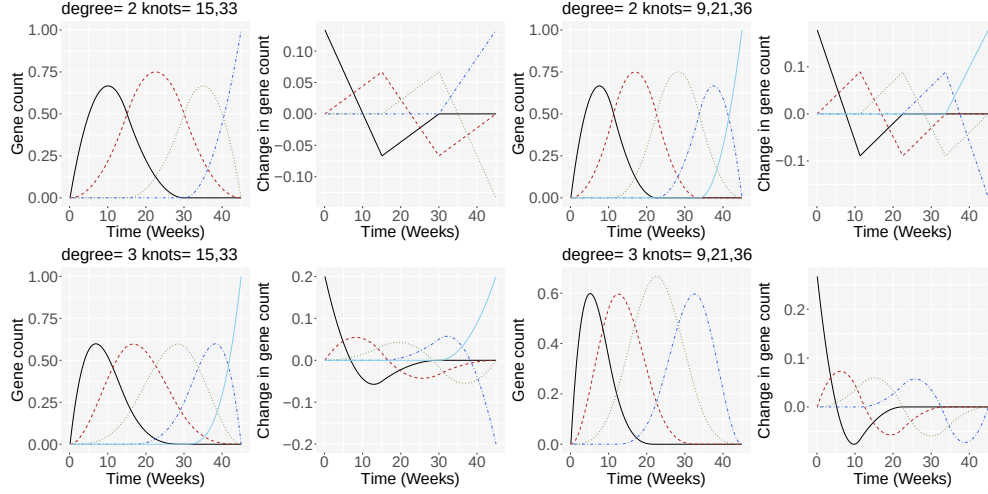


Figure A: A depiction of the different B-spline basis functions (left) and the respective first derivatives (right) used in Section 4. Of note, as per the default option `intercept=F` of the R function `bSpline()` used for the generation of the B-spline basis, the first basis function is left out (Wang, W. and Yan, J., 2021). Therefore, the number of basis functions is $d - 1$ and not d as the theoretical approach would indicate. The intercept is included in the estimation via the R function `lm()`.

data) and used to estimate the empirical $(\alpha/2)$ - and $(1 - (\alpha/2))$ -quantiles, say $q_{\alpha/2}$ and $q_{(1-(\alpha/2))}$ providing percentile confidence intervals for the time points of interest of the original data set that can be expressed as $[q_{\alpha/2}, q_{(1-(\alpha/2))}]$. Figure B depicts a visualization of this concept, while Algorithm A summarizes the procedure. For the application presented in Section 4, we chose $B_3 = 500$ bootstrap runs. This is consistent with the choice of $B_1 = 500$ for the first-level bootstrap runs, since quantiles are estimated in both cases.

4 Application to an individual gene - detailed analysis

Cd163

'Cd163' is a gene found in both mice (National Center for Biotechnology Information, 2024b) and humans (National Center for Biotechnology Information, 2024a). The soluble form of the encoded molecule was found to be a predictor for both fibrosis and hepatocellular carcinoma development in nonalcoholic steatohepatitis in humans, as reported in Kawanaka, M. and Nishino, K. and Kawada, M. and Ishii, K. and Tanikawa, T. and Katsumata, R. and Urata, N. and Nakamura, J. and Suehiro, M. and Haruma, K. and Kawamoto, H (2023). In particular, decreased levels were observed in patients with NAFLD who exhibited alleviated fibrosis and inflammation. In Ghallab, A. et al. (2021), the gene has been identified as a RJG. This implies that the gene in question is only deregulated after a specific period of time, in this case, after 24 weeks. In a preceding model selection step, the 4pLL model, the exponential model, the

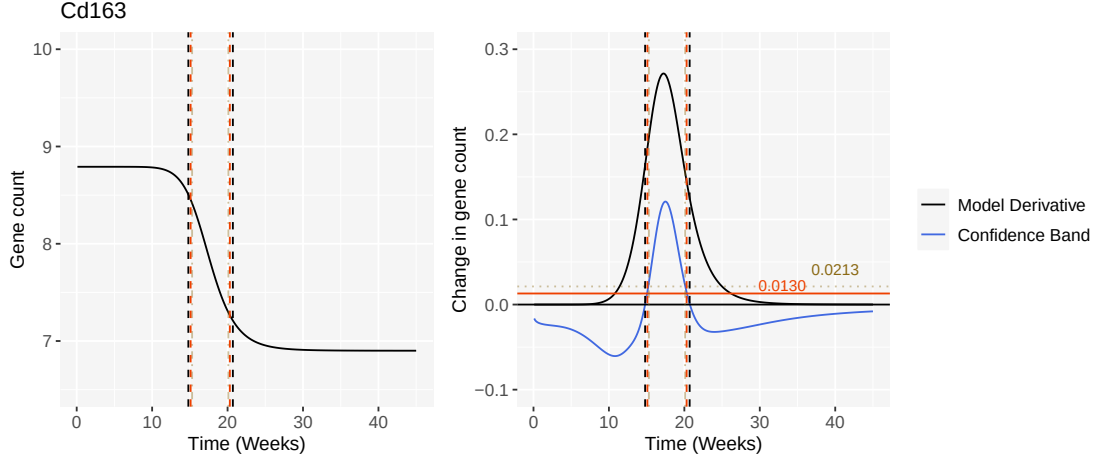


Figure B: On the left, the fit of the 4pLL model to the normalized gene count of the gene 'Cd163' over time is depicted. On the right, the first derivative and the estimated confidence band of the model fit are displayed. The horizontal red line indicates $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines indicate the corresponding time frame of significant change in gene expression. The yellow colored lines correspond to an alternative choice of $\lambda = \log_2(1.5) \times 27.5^{-1}$. The black lines represent the case $\lambda = 0$.

quadratic model and the beta model were fitted to the gene expression data of the WD-fed mice. The AIC for the 4pLL model was 90.4, whereas for the other models it was between 98.2 and 102.8. Therefore, the 4pLL model was used for further analyses.

The left panel of Figure B depicts the estimated gene expression curve. In accordance with the identification as a RJG, it appears that approximately 10 – 12 weeks pass, before the gene becomes down-regulated. Following an additional 10 – 12 weeks, the deregulated expression appears to reach a constant value. Overall, the gene changes by a value of -1.891 . In the right panel of Figure B, the corresponding first derivative and the estimated confidence band are depicted. The various horizontal lines represent different potential choices for the threshold λ , while the corresponding vertical lines represent the respective time frames of significant change in gene expression. The black line represents the case where $\lambda = 0$, while the red line corresponds to the aforementioned case $\lambda = \log_2(1.5) \times 45^{-1} = 0.0130$ (see Section 2.2). Similarly, a third threshold was calculated, indicated by the yellow line, which corresponds to a change of $\log_2(1.5)$ in gene expression over the span of half the study. This third threshold is given by $\lambda = \log_2(1.5) \times 27.5^{-1} = 0.0213$.

We observe that H_0 is rejected for all thresholds. In the following, we will present the analysis for $\lambda = \log_2(1.5) \times 45^{-1}$ as an illustrative example. The period during which the estimated significant changes in gene expression occur begins at 15.1 weeks (95% CI [13.2, 17.8]) and ends at 20.3 weeks (95% CI [17.4, 23.5]). Consequently, we can conclude the classical RJG form of a delayed deregulation. The change is observed over a period of 5.2 weeks, with a plateau following the initial 'jump' in activity. Over the course of the 5.2 week period, the expression

undergoes a change of 1.189. Therefore, we can account for 62.883% of the observed absolute maximal change, 1.891, via a significant effect. Confidence intervals for the time points of interest were estimated using a third-level bootstrap.

For comparison, the other cases indicate a time period of significant change in gene expression between weeks 14.8 to 20.7 ($\lambda = 0$, black) and 15.3 to 20.1 ($\lambda = 0.0213$, yellow), respectively. Consequently, while the 'Cd163' gene was also identified as a RJG, the estimated time frame of deregulation is set earlier than reported in Ghallab, A. et al. (2021) for all thresholds.

As described in Section 4.1. of the main manuscript, the findings derived from fitting B-splines are overall quite similar.

Fam83a

'Fam83a' is a gene found in both mice (National Center for Biotechnology Information, 2025b) and humans (National Center for Biotechnology Information, 2025a), which is associated with the progression of hepatocellular carcinoma (Yu, W.-S. and Wang, Z.-G. and Guo, R.-P. and Lin, Z.-Q. and Ye, Z.-W. and Lu, C.-L., 2020). In Ghallab, A. et al. (2021), the gene has been identified as a RJG that becomes deregulated after 18 weeks. In a preceding model selection step, the 4pLL model, the exponential model, the quadratic model and the beta model were fitted to the gene expression data of the WD-fed mice. The AIC for the beta model was 85.7, whereas for the other models it was between 86.7 and 114.7. Therefore, the beta model was used for further analyses.

The left panel of Figure C depicts the estimated gene expression curve. Contrary to the identification as a RJG, it appears that it becomes deregulated immediately. After a period of 20 – 30 weeks, the deregulated expression seems to attain a constant value. Overall, the gene changes by a value of 2.762. As for the example above, the right panel of Figure C depicts the corresponding first derivative, the estimated confidence band and the respective time frame of significant change in gene expression.

We observe that H_0 is rejected for all thresholds. In the following, we will present the analysis for $\lambda = \log_2(1.5) \times 45^{-1}$ as an illustrative example. The period during which the estimated significant changes in gene expression occur begins at 0.3 weeks (95% CI [0.1, 1.9]) and ends at 24.9 weeks (95% CI [15.7, 32.1]). The change is observed over a period of 24.6 weeks. Over the course of this period, the expression undergoes a change of 2.380. Therefore, we can account for almost all of the observed absolute maximal change via a significant effect.

For comparison, the other cases indicate a time period of significant change in gene expression between weeks 0.3 to 20.7 ($\lambda = 0$, black) and 0.3 to 18.6 ($\lambda = 0.0213$, yellow), respectively. Consequently, while the 'Fam83a' gene was identified as a RJG in Ghallab, A. et al. (2021), the estimated time frame of deregulation begins with the start of the study for all thresholds.

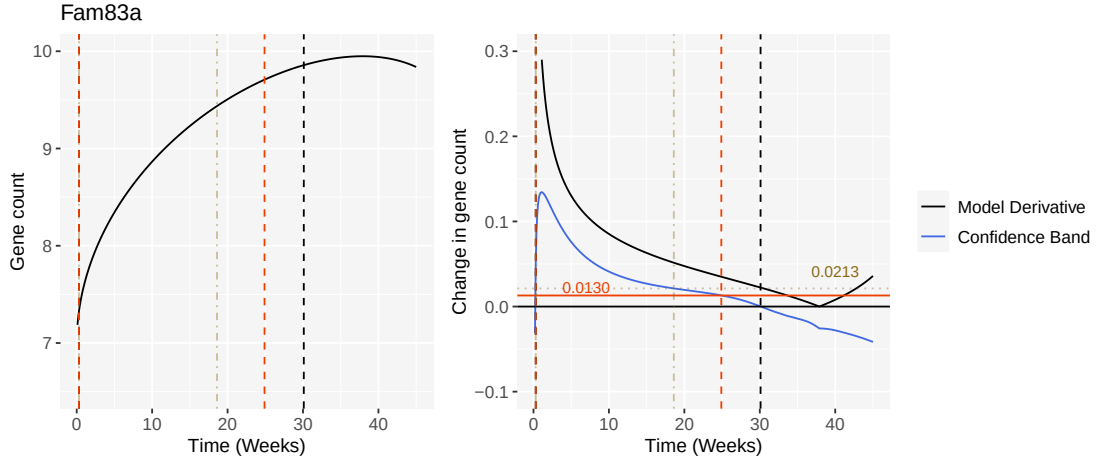


Figure C: On the left, the fit of the beta model to the normalized gene count of the gene 'Fam83a' over time is depicted. On the right, the first derivative and the estimated confidence band of the model fit are displayed. The horizontal red line indicates $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines indicate the corresponding time frame of significant change in gene expression. The yellow lines correspond to an alternative choice of $\lambda = \log_2(1.5) \times 27.5^{-1}$. The black lines represent the case $\lambda = 0$.

Additionally, an analysis based on fitting B-splines to the data was conducted. The procedure was identical to the one described for the 'Cd163' gene in Section 4.1 of the main manuscript. The results are depicted in Figure D. Overall, significant change in gene expression was detected from 4 weeks until 17.5–22.6 weeks, which is slightly later but mostly consistent with the results obtained by applying the parametric approach. An exception is observed in the case of $l = 3$ with three knots, wherein the first estimated time period starts with the beginning of the study. However, the fit to the data seems to indicate overfitting. For all cases, except for $l = 2$ degrees and two knots, the time frame of significant change is segmented by an inactive period around week 10. These results are caused by the notable spike in expression at 6 weeks on the study or equivalently a dip of the measurements taken at 9 weeks on the study. This may be a fragment of the data that was incorrectly identified as a period of inactivity due to the flexible nature of B-splines. Nevertheless, it is conceivable that an inactive period exists, which would not be discernible through the application of a comparatively rigid parametric model.

Dbp

'Dbp' is a protein encoding gene detected in humans (National Center for Biotechnology Information, 2024c) and mice (National Center for Biotechnology Information, 2024d) associated with the circadian rhythm. In a preceding model selection step, the 4pLL model, the exponential model, the quadratic model and the beta model were fitted to the gene expression data of the WD-fed mice. The AIC for the exponential model was 141.9, whereas for the other models it

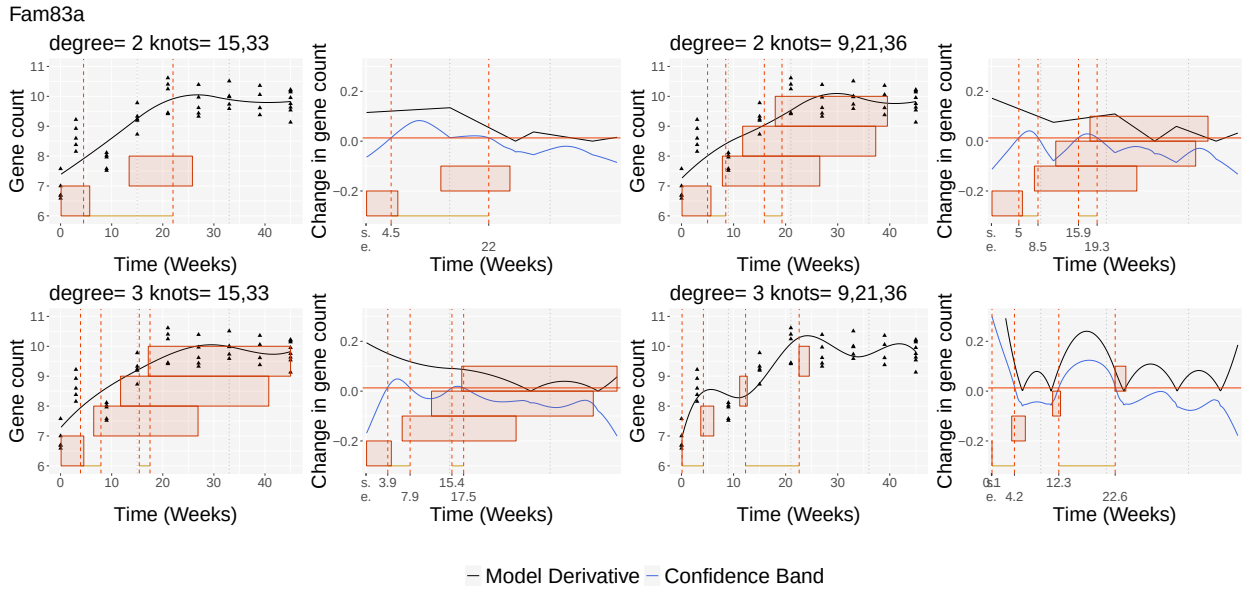


Figure D: The fit of the B-splines to the normalized gene count of the gene 'Fam83a' over time is depicted on the left for each presented combination of degree and number of knots. The degree of the basis functions and the knot placement are stated in the subtitle of the respective plots. Additionally, the knots are indicated by gray dotted lines. On the right, the first derivatives and the estimated confidence bands are displayed. The horizontal red lines indicate $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines and the horizontal yellow lines indicate the corresponding time frames of significant change in gene expression. The red boxes illustrate the estimated confidence intervals for the respective time points of interest.

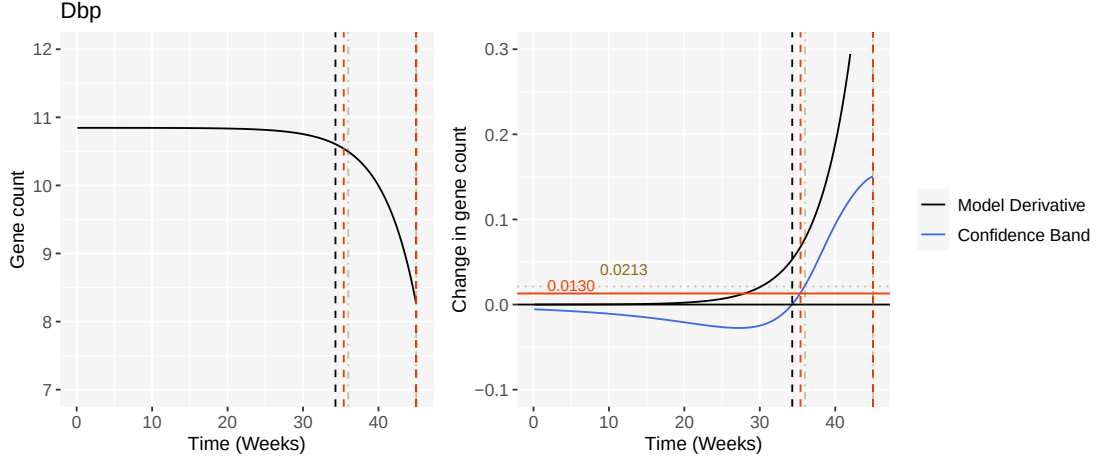


Figure E: On the left, the fit of the exponential model to the normalized gene count of the gene 'Dbp' over time is depicted. On the right, the first derivative and the estimated confidence band of the model fit are displayed. The horizontal red line indicates $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines indicate the corresponding time frame of significant change in gene expression. The yellow lines correspond to an alternative choice of $\lambda = \log_2(1.5) \times 27.5^{-1}$. The black lines represent the case $\lambda = 0$.

was between 143.8 and 148.4. Therefore, the exponential model was used for further analyses. The left panel of Figure E depicts the estimated gene expression curve. It appears that it becomes deregulated after approximately 30 weeks. Overall, the gene changes by a value of -2.580 . As for the example above, the right panel of Figure E depicts the corresponding first derivative, the estimated confidence band and the respective time frames of significant change in gene expression.

We observe that H_0 is rejected for all thresholds. In the following, we will present the analysis for $\lambda = \log_2(1.5) \times 45^{-1}$ as an illustrative example. The period during which the estimated significant changes in gene expression occur begins at 35.4 weeks (95% CI [29.3, 37.0]) and appears to continue after the end of the study. The change is observed over a period of 9.6 weeks. Over the course of this period, the expression undergoes a change of 2.275. Therefore, we can account for more than 85% of the observed absolute maximal change, 2.580, via a significant effect.

For comparison, the other cases indicate a time period of significant change in gene expression between weeks 34.3 ($\lambda = 0$, black) and 36.0 ($\lambda = 0.0213$, yellow) till the end of the study, respectively. Consequently, we can identify the 'Dbp' gene as a RJG.

Additionally, an analysis based on fitting B-splines to the data was conducted. The procedure was identical to the one described for the 'Cd163' gene in Section 4.1. of the main manuscript. The results are depicted in Figure F. Overall, two distinct time frames of significant change in expression were detected. The beginning of the later period was detected at 35.5 – 39.7

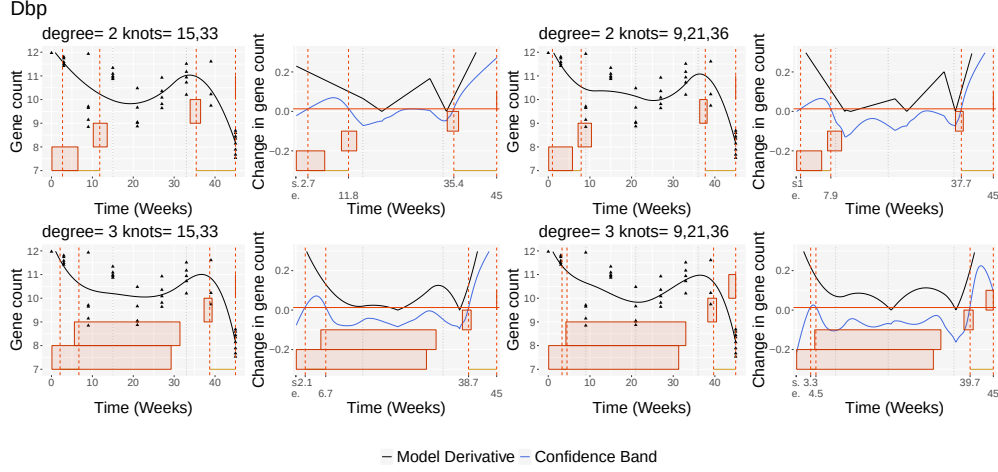


Figure F: The fit of the B-splines to the normalized gene count of the gene 'Dbp' over time is depicted on the left for each presented combination of degree and number of knots. The degree of the basis functions and the knot placement are stated in the subtitle of the respective plots. Additionally, the knots are indicated by gray dotted lines. On the right, the first derivatives and the estimated confidence bands are displayed. The horizontal red lines indicate $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines and the horizontal yellow lines indicate the corresponding time frames of significant change in gene expression. The red boxes illustrate the estimated confidence intervals for the respective time points of interest.

weeks up until the end of the study, which is consistent with the results obtained applying the parametric approach. However, the early period of activity was not detected using the parametric model. For $l = 2$ degrees, this period is estimated to last 6.9 – 9.1 weeks from 1 – 2.7 weeks until 7.9 – 11.8 weeks. For $l = 3$ degrees, the period is shorter. It is estimated to start at 2.1 – 3.3 weeks until 4.5 – 6.7 weeks. These results demonstrate the benefit of fitting B-splines, since the earlier period of activity was most likely missed by the parametric model due to the predefined shape.

Tm7sf2

'Tm7sf2' is a gene found in both mice (National Center for Biotechnology Information, 2025c) and humans (National Center for Biotechnology Information, 2025d), which is associated with weight gain in mice (Gatticchi, L. and de Las Heras, J.I. and Sivakumar, A. and Zuleger, N. and Roberti, R. and Schirmer, E.C., 2020). In a preceding model selection step, the 4pLL model, the exponential model, the quadratic model and the beta model were fitted to the gene expression data of the WD-fed mice. The AIC for the quadratic model was 32.0, whereas for the other models it was between 32.8 and 40.9. Therefore, the quadratic model was used for further analyses.

The left panel of Figure G depicts the estimated gene expression curve. It appears that it

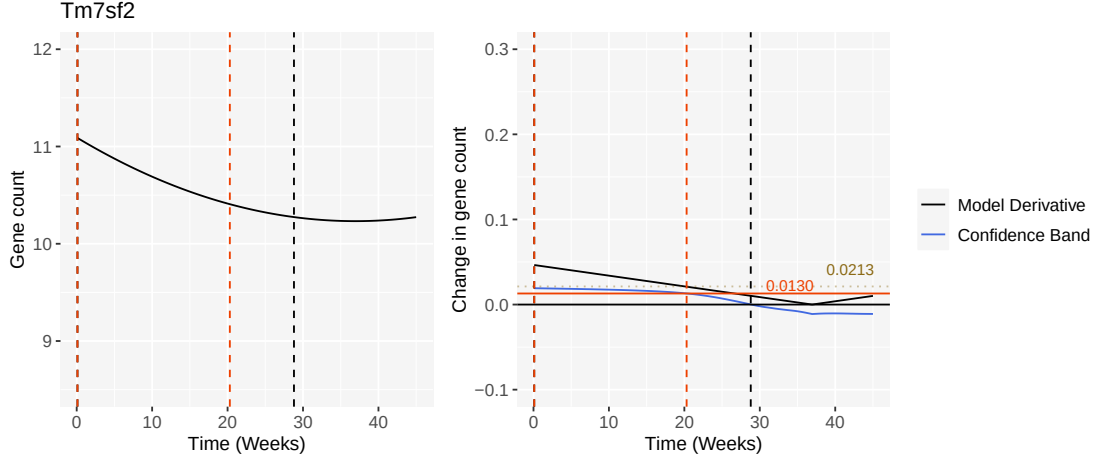


Figure G: On the left, the fit of the quadratic model to the normalized gene count of the gene 'Tm7sf2' over time is depicted. On the right, the first derivative and the estimated confidence band of the model fit are displayed. The horizontal red line indicates $\lambda = \log_2(1.5) \times 45^{-1}$. The vertical dashed red lines indicate the corresponding time frame of significant change in gene expression. The yellow colored lines correspond to an alternative choice of $\lambda = \log_2(1.5) \times 27.5^{-1}$. The black lines represent the case $\lambda = 0$.

becomes deregulated immediately after the start of the study and reaches a plateau after week 30. Overall the gene changes by a value of -0.855 . As for the example above, the right panel of Figure G depicts the corresponding first derivative, the estimated confidence band and the respective time frames of significant change in gene expression.

We observe that H_0 is rejected only for the thresholds $\lambda = 0$ and $\lambda = \log_2(1.5) \times 45^{-1}$. H_0 can not be rejected to the α -level 0.05 for $\lambda = \log_2(1.5) \times 27.5^{-1} = 0.0213$. In the following, we will present the analysis for $\lambda = \log_2(1.5) \times 45^{-1}$ as an illustrative example. The period during which the estimated significant changes in gene expression occur begins at the start of the study (95% CI [0.1, 14.0]) and end at week 20.3 (95% CI [12.2, 25.6]). The change is observed over a period of 20.2 weeks. Over the course of this period, the expression undergoes a change of 0.681. Therefore, we can account for almost 80% of the observed absolute maximal change, 0.855, via a significant effect.

For comparison, the case $\lambda = 0$ indicates a time period of significant change in gene expression between the beginning of the study and week 28.8.

Additionally, an analysis based on fitting B-splines to the data was conducted. The procedure was identical to the one described for the 'Cd163' gene in Section 4.1. of the main manuscript. The results are depicted in Figure H. Interestingly, there is no significant change in expression detected. We note that the signal that results in the detection of the significant change in expression applying the quadratic model is very small. These results demonstrate the benefit of fitting a parametric model, since the procedure, given a well specified model, appears to be

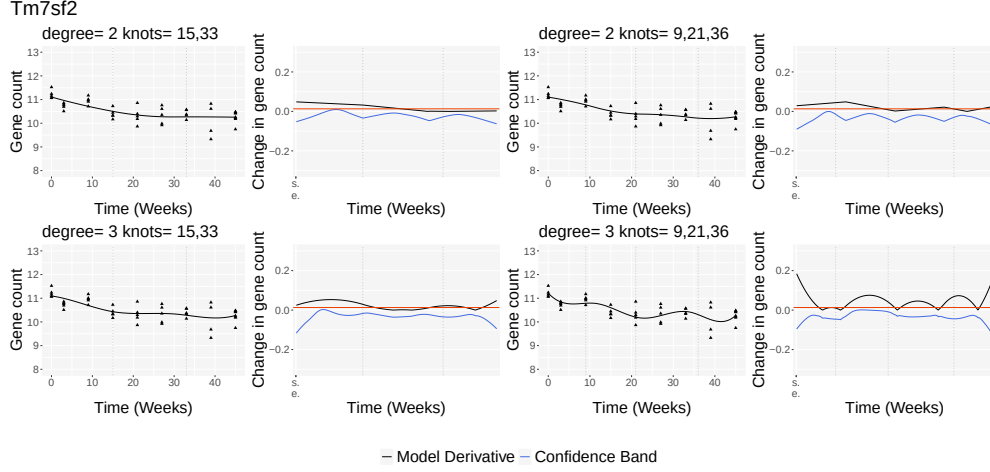


Figure H: The fit of the B-splines to the normalized gene count of the gene 'Tm7sf2' over time is depicted on the left for each presented combination of degree and number of knots. The degree of the basis functions and the knot placement are stated in the subtitle of the respective plots. Additionally, the knots are indicated by gray dotted lines. On the right, the first derivatives and the estimated confidence bands are displayed. The horizontal red lines indicates $\lambda = \log_2(1.5) \times 45^{-1}$.

more sensitive.

5 Comparison of two groups

For this application we consider a gene that deregulates in both diet groups and compare the estimated time frames of significant change in gene expression. The biological question inspiring this example is whether the WD accelerates or decelerates, and thereby 'ages' biological processes, which also occur in SD-fed mice. The chosen example, 'Dbp', is a protein encoding gene detected in humans (National Center for Biotechnology Information, 2024c) and mice (National Center for Biotechnology Information, 2024d) associated with the circadian rhythm. A preliminary model selection step indicated a superior fit of the exponential model to the gene expression data for both diet groups. Consequently, it was utilised in the subsequent analysis.

The left panel of Figure I depicts the modeled gene expression of both the SD and the WD groups. The gene count in SD-fed mice initially appears to be smaller and both cases seem to undergo similar deregulation at approximately the same time and converge to the same value. On the right, the corresponding first derivatives and the estimated confidence bands are displayed. The dotted line represents the time period of significant change in gene expression for the WD-fed group, while the dot-dashed line outlines the time period for the SD-fed mice. Although the first derivative of the WD-fed group appears to increase earlier, the confidence band of the SD group exceeds the threshold $\lambda = \log_2(1.5) \times 45^{-1} = 0.0130$ earlier. For a

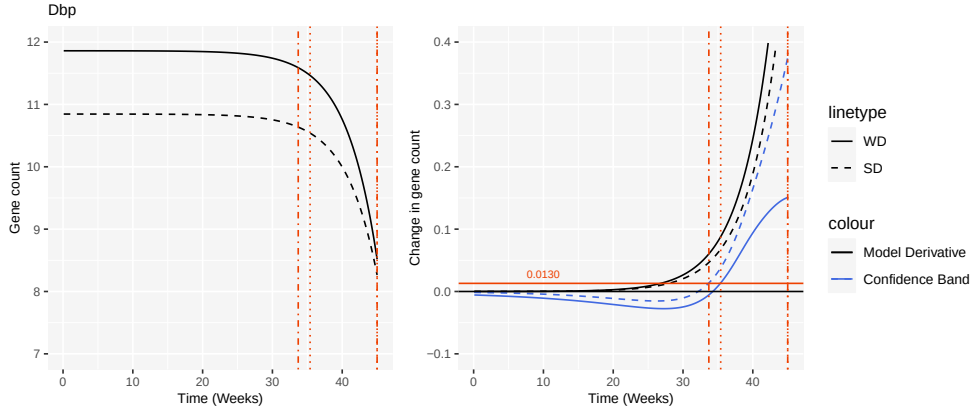


Figure I: A comparison of the SD versus the WD, exemplary for the gene 'Dbp'. The gene counts, as modeled by the exponential model, are depicted on the left. On the right, the first derivatives and their estimated confidence bands are shown. The red dotted line outlines the time period of significant change in gene expression for the WD, while the dot-dashed line outlines the analogous period for the SD.

detailed discussion of the results for the WD-fed mice see the section above. The estimated time period during which significant change in gene expression in SD-fed mice occurs begins at 33.7 weeks (95% CI [28.8, 34.4]) and ends with the end of study at 45 weeks (95% CI [45.0, 45.0], for computational reasons). The gene count undergoes a change of 3.070 over the span of 11.3 weeks. Overall, both genes show very similar results. The estimated confidence band for the SD-fed mice exceeds the threshold $\lambda = 0.0130$ about -1.7 weeks earlier and the estimated 95% confidence interval for this difference is given by $[-6.4, 4.1]$.

It should be noted that no measurements were taken from SD-fed mice between weeks 3 and 27. Hence, when interpreting the results of a comparison of the two diets during this time period, this should be taken into account. Since the parametric fit indicates activity only after week 30, this is not a problem for the results presented here. However, due to their flexible nature, this causes problems when applying the method using B-splines. Therefore, we limit the proof of concept presented here to the parametric approach.

References

- Bornkamp, B. and Pinheiro, J. and Bretz, F. MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies. *Journal of Statistical Software*, 29(7):1 – 23, 2009. doi: 10.18637/jss.v029.i07.
- Bornkamp, B. and Pinheiro, J. and Bretz, F. *MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies*, 2017. URL <https://CRAN.R-project.org/package=>

MCPMod. R package version 1.0-10.1.

- Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B.D. *Regression: Models, Methods and Applications*. Springer Berlin, Heidelberg, 2nd edition, 2022. doi: 10.1007/978-3-662-63882-8.
- Gatticchi, L. and de Las Heras, J.I. and Sivakumar, A. and Zuleger, N. and Roberti, R. and Schirmer, E.C. Tm7sf2 disruption alters radial gene positioning in mouse liver leading to metabolic defects and diabetes characteristics. *Frontiers in Cell and Developmental Biology*, 8:592573, 2020. doi: 10.3389/fcell.2020.592573.
- Ghallab, A., Myllys, M., Friebe, A., Duda, J., Edlund, K., Halilbasic, E., Vucur, M., Hobloss, Z., Brackhagen, L., Begher-Tibbe, B., Hassan, R., Burke, M., Genc, E., Frohwein, L.J., Hofmann, U., Holland, C.H., González, D., Keller, M., Seddek, A., Abbas, T., Mohammed, E.S.I., Teufel, A., Itzel, T., Metzler, S., Marchan, R., Cadenas, C., Watzl, C., Nitsche, M.A., Kappenberg, F., Luedde, T., Longerich, T., Rahnenführer, J., Hoehme, S., Trauner, M., and Hengstler, J.G. Spatio-Temporal Multiscale Analysis of Western Diet-Fed Mice Reveals a Translationally Relevant Sequence of Events during NAFLD Progression. *Cells*, 10(10), 2021. doi: 10.3390/cells10102516.
- Kawanaka, M. and Nishino, K. and Kawada, M. and Ishii, K. and Tanikawa, T. and Katsumata, R. and Urata, N. and Nakamura, J. and Suehiro, M. and Haruma, K. and Kawamoto, H. Soluble CD163 is a predictor of fibrosis and hepatocellular carcinoma development in nonalcoholic steatohepatitis. *BMC gastroenterology*, 23(1), 2023. doi: 10.1186/s12876-023-02786-4.
- National Center for Biotechnology Information. *CD163 molecule [Homo sapiens (human)]*. 2024a. URL <https://www.ncbi.nlm.nih.gov/gene/93671>. Accessed June 13, 2024.
- National Center for Biotechnology Information. *Cd163 antigen [Mus musculus (house mouse)]*. 2024b. URL <https://www.ncbi.nlm.nih.gov/gene/93671>. Accessed June 13, 2024.
- National Center for Biotechnology Information. *DBP D-box binding PAR bZIP transcription factor [Homo sapiens (human)]*. 2024c. URL <https://www.ncbi.nlm.nih.gov/gene/1628>. Accessed June 14, 2024.
- National Center for Biotechnology Information. *Dbp D site albumin promoter binding protein [Mus musculus (house mouse)]*. 2024d. URL <https://www.ncbi.nlm.nih.gov/gene/13170>. Accessed June 14, 2024.
- National Center for Biotechnology Information. *FAM83A family with sequence similarity 83 member A [Homo sapiens (human)]*. 2025a. URL <https://www.ncbi.nlm.nih.gov/gene/84985>. Accessed May 16, 2025.

- National Center for Biotechnology Information. *Fam83a family with sequence similarity 83, member A [Mus musculus (house mouse)]*. 2025b. URL <https://www.ncbi.nlm.nih.gov/gene/239463>. Accessed May 16, 2025.
- National Center for Biotechnology Information. *Tm7sf2 transmembrane 7 superfamily member 2 [Mus musculus (house mouse)]*. 2025c. URL <https://www.ncbi.nlm.nih.gov/gene/73166>. Accessed May 16, 2025.
- National Center for Biotechnology Information. *TM7SF2 transmembrane 7 superfamily member 2 [Homo sapiens (human)]*. 2025d. URL <https://www.ncbi.nlm.nih.gov/gene/7108>. Accessed May 16, 2025.
- Wang, W. and Yan, J. Shape-restricted regression splines with R package splines2. *Journal of Data Science*, 19(3):498–517, 2021. doi: 10.6339/21-JDS1020.
- Yu, W.-S. and Wang, Z.-G. and Guo, R.-P. and Lin, Z.-Q. and Ye, Z.-W. and Lu, C.-L. Hepatocellular carcinoma progression is protected by miRNA-34c-5p by regulating FAM83A. *European Review for Medical & Pharmacological Sciences*, 24(11), 2020. doi: 10.26355/eurev_202006_21499.

Table A: Levels of standard deviation used for the simulation study. The standard deviations marked with an asterisk (*) are approximately half the size of the maximum change in gene expression.

Scce.	small	mid-small	medium	mid-large	large
1	0.014	0.021	0.028	0.042	0.056*
2	0.149	0.223	0.297	0.446	0.595*
3	0.595	0.892	1.190	1.784*	2.380*
4	0.283	0.424	0.566	0.849	1.132
5	0.283	0.424	0.566	0.849	1.132
6	0.283	0.424	0.566	0.849	1.132

Table B: Simulation results. 'CorRuns' indicates the number of simulation runs, for which the correct number of coherent subsets of $\tilde{S}$ were detected. The suffix '_s1/2' indicates the start of $S_{1/2}$ and correspondingly '_e1' the end. Since S_2 ends at the end of study, it is not considered here.

Sci.	σ	CorRuns	Bias_s1	Var_s1	Bias_e1	Var_e1	Bias_s2	Var_s2
Sce. 2	small σ	998	2.339	0.584	-2.819	0.917	-	-
Sce. 2	mid-small σ	935	3.323	1.319	-3.974	1.896	-	-
Sce. 2	medium σ	828	3.877	1.849	-4.745	2.424	-	-
Sce. 2	mid-large σ	625	4.823	2.777	-5.16	3.893	-	-
Sce. 2	large σ	371	5.413	6.359	-5.24	8.592	-	-
Sce. 3	small σ	989	5.398	2.93	-10.135	8.996	-	-
Sce. 3	mid-small σ	617	7.056	7.206	-13.21	20.138	-	-
Sce. 3	medium σ	442	8.431	15.564	-14.038	35.942	-	-
Sce. 3	mid-large σ	392	10.421	10.989	-16.256	16.88	-	-
Sce. 3	large σ	301	11.369	10.746	-16.347	16.157	-	-
Sce. 4	small σ	118	0.005	0.001	-3.951	1.85	2.631	1.001
Sce. 4	mid-small σ	44	0.12	0.046	-5.963	5.435	2.4	1.627
Sce. 4	medium σ	43	0.358	0.285	-8.897	17.891	2.188	2.042
Sce. 4	mid-large σ	33	1.27	3.535	-16.938	34.394	2.076	2.706
Sce. 4	large σ	37	2.379	10.036	-21.142	35.308	1.882	3.45
Sce. 5	small σ	999	3.956	0.981	-1.225	0.236	2.074	1.065
Sce. 5	mid-small σ	864	5.267	1.375	-1.845	0.567	2.912	2.053
Sce. 5	medium σ	671	6.244	1.912	-2.535	1.174	3.141	1.999
Sce. 5	mid-large σ	424	7.5	4.88	-4.932	11.235	3.475	1.909
Sce. 5	large σ	260	8.865	11.986	-7.61	27.344	3.339	12.426
Sce. 6	small σ	993	0	0	-6.996	7.686	-	-
Sce. 6	mid-small σ	992	0	0	-11.503	22.59	-	-
Sce. 6	medium σ	995	0.003	0.001	-16.461	33.71	-	-
Sce. 6	mid-large σ	964	0.103	0.153	-21.628	27.636	-	-
Sce. 6	large σ	749	0.375	1.405	-23.591	21.108	-	-

Table C: Top 20 results of the GO analysis performed on the selected 569 genes. An. for Annotated; Sign. for Significant; Exp. for Expected

Term	An.	Sign.	Exp.	go_pval_adj
alpha-amino acid catabolic process	57	16	3.33	0.0005
surfactant homeostasis	13	8	0.76	0.0005
positive regulation of phagocytosis	70	17	4.09	0.0009
phagocytosis	175	36	10.23	0.0025
positive regulation of tumor necrosis fa...	97	19	5.67	0.0035
branched-chain amino acid catabolic proc...	13	7	0.76	0.0035
negative regulation of interleukin-6 pro...	39	11	2.28	0.0099
negative regulation of IP-10 production	7	5	0.41	0.0106
xenobiotic metabolic process	72	15	4.21	0.0106
positive regulation of interleukin-6 pro...	81	16	4.73	0.0106
bile acid biosynthetic process	23	8	1.34	0.0199
negative regulation of cytokine producti...	18	7	1.05	0.0252
regulation of B cell differentiation	32	9	1.87	0.0361
regulation of ribonuclease activity	9	5	0.53	0.0373
amino acid catabolic process	70	19	4.09	0.0383
positive regulation of interleukin-10 pr...	33	9	1.93	0.0383
inflammatory response	525	64	30.68	0.0445
positive regulation of adaptive immune r...	91	18	5.32	0.0478
mast cell activation	55	14	3.21	0.0478
immunoglobulin mediated immune response	133	23	7.77	0.0492

Algorithm A. Let $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Sigma}}$, $\tilde{S}$ and the time points of interest be calculated from the original data set as described in Section 2. In the event of independent measurements, it is possible to replace $\hat{\boldsymbol{\Sigma}}$ with $\hat{\sigma}^2$, as outlined in Section 3.

Step 1 Generate bootstrap data at the m different time points using $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$. Do so by drawing iid. error terms $\boldsymbol{\varepsilon}_p^* = (\varepsilon_{p,1}^*, \dots, \varepsilon_{p,m}^*)^* \sim N_m(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$ and applying the formula

$$y_{p,q}^* = f(t_q, \hat{\boldsymbol{\theta}}) + \varepsilon_{p,q}^* \text{ for } p = 1, \dots, n_q, \quad q = 1, \dots, m.$$

If the measurements are independent: Draw iid. errors $\varepsilon_{p,q}^* \sim N(0, \hat{\sigma}^2)$.

Repeat this step B_3 times. This returns $y_{p,q}^{*,1}, \dots, y_{p,q}^{*,B_3}$.

Next, calculate the estimates $\hat{\boldsymbol{\theta}}_k^*$ and $\hat{\boldsymbol{\Sigma}}_k^*$ (or $\hat{\sigma}_k^{2,*}$), $k = 1, \dots, B_3$ as described in Section 2.1. for each of the generated data sets. This yields $|f'(t, \hat{\boldsymbol{\theta}}_k^*)|$.

Now for $k = 1, \dots, B_3$ perform the two-step bootstrap procedure to estimate the confidence band:

Step 2 For $k = 1, \dots, B_3$: Generate bootstrap data at the m different time points using $\hat{\boldsymbol{\theta}}_k^*$ and $\hat{\boldsymbol{\Sigma}}_k^*$. Do so by drawing iid. error terms $\boldsymbol{\varepsilon}_p^{**} = (\varepsilon_{p,1}^{**}, \dots, \varepsilon_{p,m}^{**}) \sim N_m(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_k^*)$ and applying the formula

$$y_{p,q}^{**} = f(t_q, \hat{\boldsymbol{\theta}}_k^*) + \varepsilon_{p,q}^{**} \text{ for } p = 1, \dots, n_q, \quad q = 1, \dots, m.$$

If the measurements are interdependent Draw iid. errors $\varepsilon_{p,q}^{**} \sim N(0, \hat{\sigma}_k^{2,*})$

Repeat this step B_1 times. This returns $y_{p,q}^{**,1}, \dots, y_{p,q}^{**,B_1}$.

For each combination of $k = 1, \dots, B_3$ and $l = 1, \dots, B_1$ estimate $\hat{\boldsymbol{\theta}}_{k,l}^{**}$ and $\hat{\boldsymbol{\Sigma}}_{k,l}^{**}$ (or $\hat{\sigma}_{k,l}^{2,**}$). This yields $|f'(t, \hat{\boldsymbol{\theta}}_{k,l}^{**})|$.

The replicates $|f'(t, \hat{\boldsymbol{\theta}}_{k,1}^{**})|, \dots, |f'(t, \hat{\boldsymbol{\theta}}_{k,B_1}^{**})|$ can be interpreted as the second level of the bootstrap. The standard error of this sample yield the estimate $\hat{\sigma}_{|f'(t, \hat{\boldsymbol{\theta}}_k^*)|}$.

Step 3 For each combination of $k = 1, \dots, B_3$ and $l = 1, \dots, B_1$: Generate bootstrap data at the m different time points using $\hat{\boldsymbol{\theta}}_{k,l}^{**}$ and $\hat{\boldsymbol{\Sigma}}_{k,l}^{**}$. Do so by drawing iid. error terms $\boldsymbol{\varepsilon}_p^{***} = (\varepsilon_{p,1}^{***}, \dots, \varepsilon_{p,m}^{***}) \sim N_m(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{k,l}^{**})$ and applying the formula

$$y_{p,q}^{***} = f(t_q, \hat{\boldsymbol{\theta}}_{k,l}^{**}) + \varepsilon_{p,q}^{***} \text{ for } p = 1, \dots, n_q, \quad q = 1, \dots, m.$$

If the measurements are independent: Draw iid. errors $\varepsilon_{p,q}^{***} \sim N(0, \hat{\sigma}_{k,l}^{2,**})$

Repeat this step B_2 times.

Estimate $\hat{\boldsymbol{\theta}}^{***}$ (indices omitted). This yields B_2 replicates of the form $|f'(t, \hat{\boldsymbol{\theta}}^{***})|$. The standard errors of the B_2 replicates yield estimates $\hat{\sigma}_{|f'(t, \hat{\boldsymbol{\theta}}_{k,l}^{**})|}$.

Step 4 For each combination of $k = 1, \dots, B_3$ and $l = 1, \dots, B_1$: Calculate

$$D^{*,k,l} := \max_{t \in T} \frac{|f'(t, \hat{\theta}_{k,l}^{**})| - |f'(t, \hat{\theta}_k^*)|}{\hat{\sigma}_{|f'(t, \hat{\theta}_{k,l}^{**})|}}.$$

The respective empiric $(1 - \alpha)$ -quantiles of the distribution yield estimates of the critical values c_k^* .

Step 5 For $k = 1, \dots, B_3$: Calculate the lower simultaneous confidence band

$$L^\alpha(t, \hat{\theta}_k^*) = |f'(t, \hat{\theta}_k^*)| + c_k^* \hat{\sigma}_{|f'(t, \hat{\theta}_k^*)|}.$$

Calculate the time frames of significant change in gene expression:

$$\tilde{S}_k := \left\{ t : L^\alpha(t, \hat{\theta}_k^*) > \lambda \right\}$$

Partition the $\tilde{S}_k$ into coherent subsets:

$$\tilde{S}_k = (S_{k,1}, \dots, S_{k,s})$$

Step 6 For $k = 1, \dots, B_3$: Dismiss the cases with more or less coherent subsets then the original data set. Identify the time points of interest for the remaining runs. Now, calculate the empiric $(\alpha/2)$ - and $(1 - (\alpha/2))$ -quantiles ($Q_{\alpha/2}$ and $Q_{(1-\alpha/2)}$) for each time point of interest, respectively. The resulting confidence intervals are:

$$\left[Q_{\frac{\alpha}{2}}, Q_{(1-\frac{\alpha}{2})} \right]$$

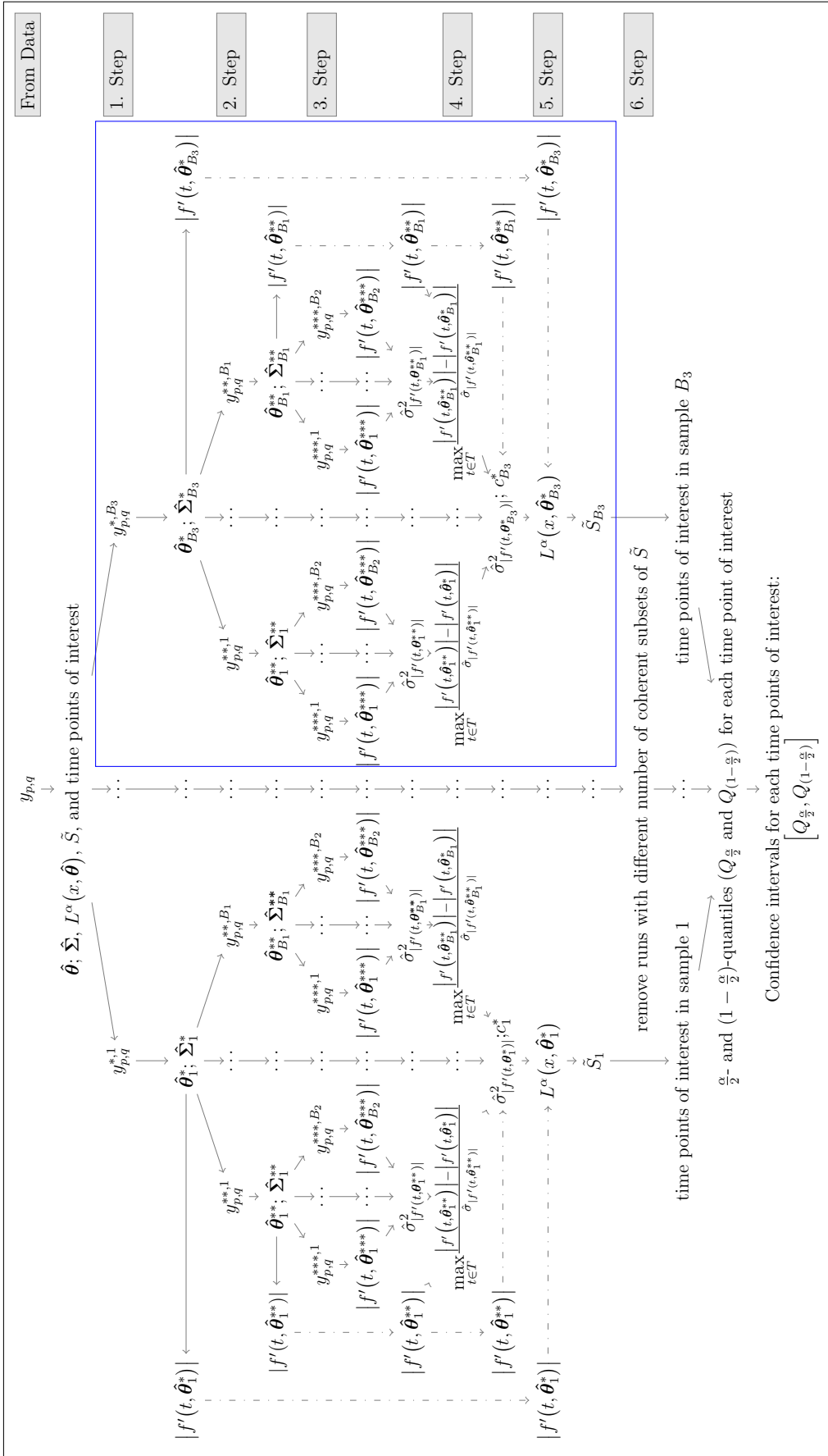


Figure J: Flowchart of the algorithm described in Subsection 2.4. The blue rectangle indicates the procedure described in Section 2.3. If all measurements are independent the assumption of independent multivariate normal distributed error vectors can be simplified to assuming independent normally distributed errors and $\hat{\Sigma}^{*,*,*,*}$ is replaced by $\hat{\sigma}^2, *,*,*,*$ as described in Section 3.1.

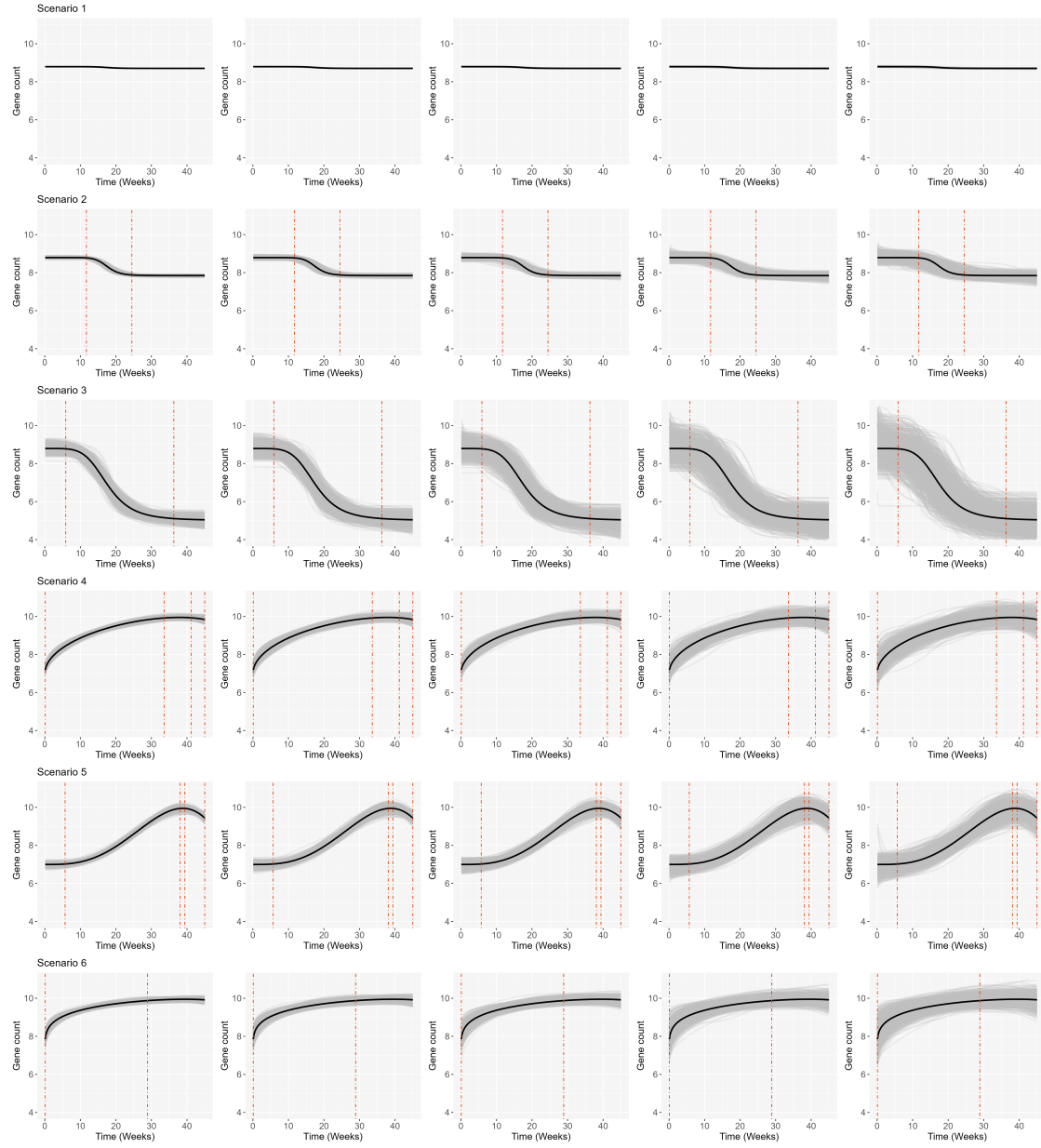


Figure K: Comparison of the true model function with the estimated model functions derived from each of the simulation studies. The standard deviation increases from left to right.

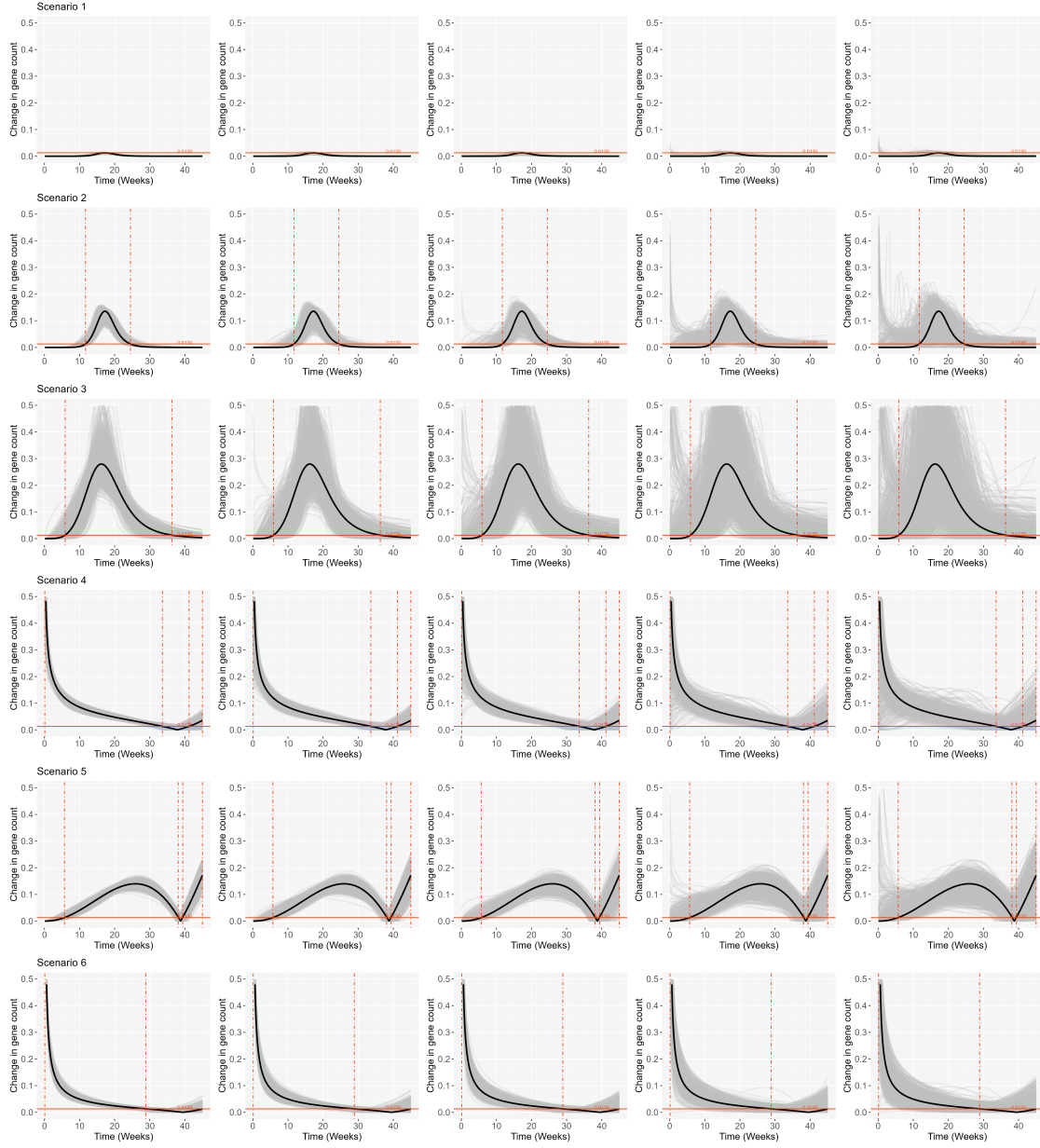


Figure L: Comparison of the true first derivative with the estimated first derivatives derived from each of the simulation studies. The standard deviation increases from left to right.

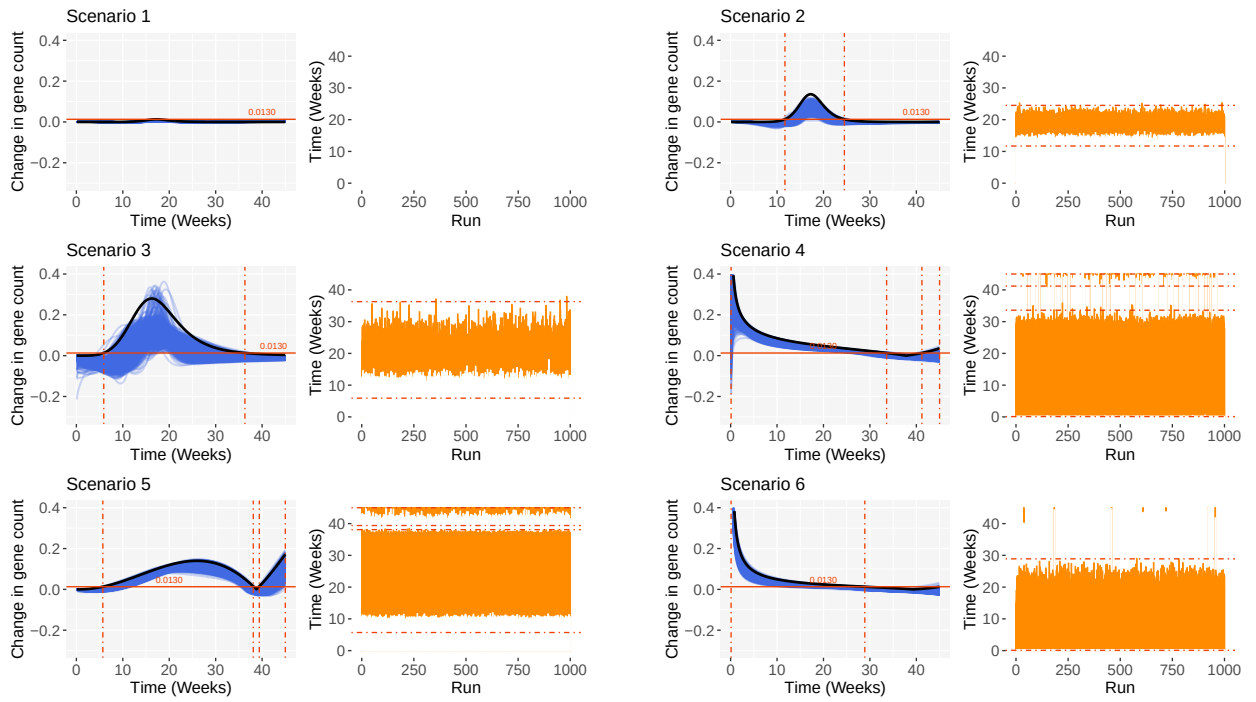


Figure M: Results of the six simulation scenarios for a small standard deviation. The left column depicts the true first derivative and the estimated confidence bands. The right column depicts the estimated time frames of significant change in gene expression for each run.

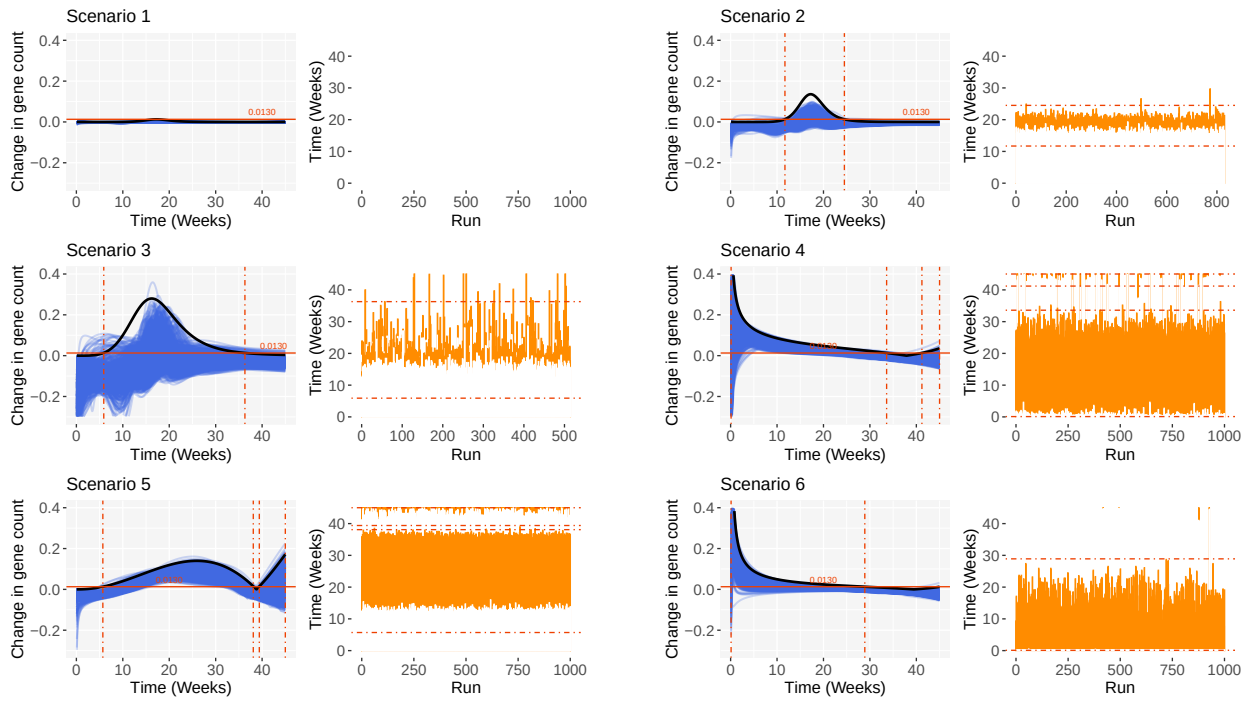


Figure N: Results of the six simulation scenarios for a medium standard deviation. The left column depicts the true first derivative and the estimated confidence bands. The right column depicts the estimated time frames of significant change in gene expression for each run.

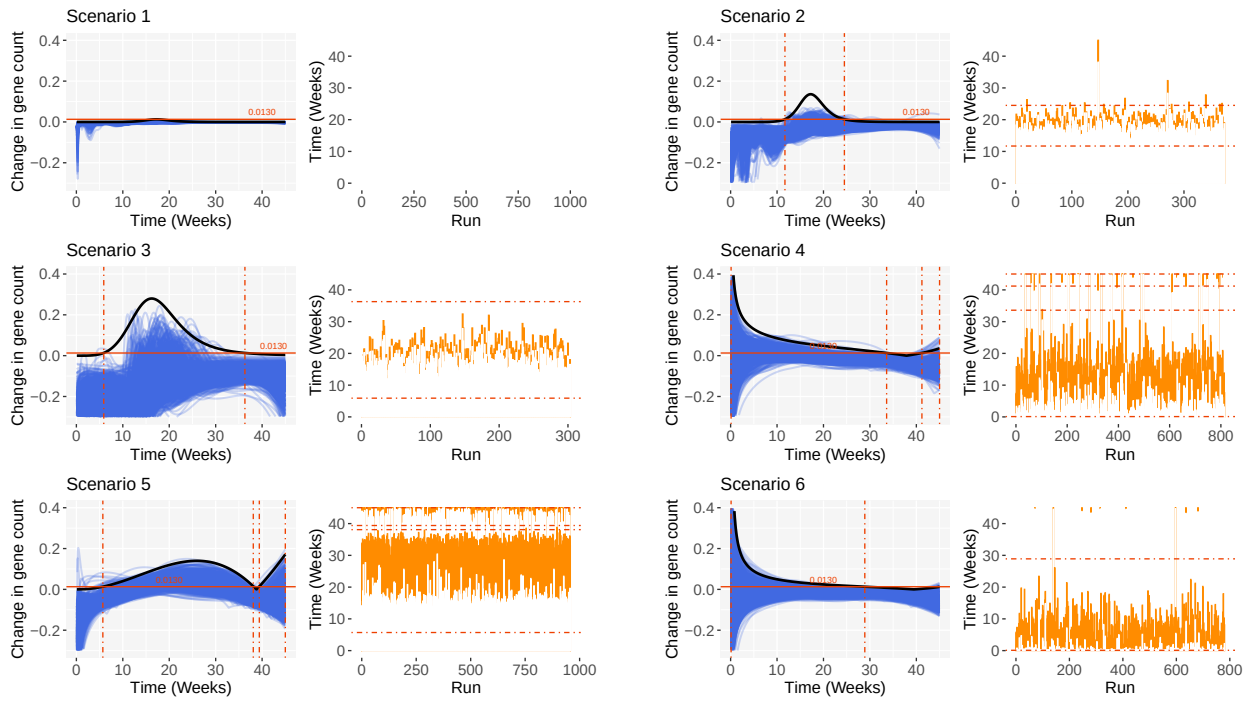


Figure O: Results of the six simulation scenarios for a large standard deviation. The left column depicts the true first derivative and the estimated confidence bands. The right column depicts the estimated time frames of significant change in gene expression for each run.

3.3 Estimating effect thresholds and beyond: A flexible framework for multivariate alert detection

In this article, the discussion of model-based alert identification is extended to two or more continuous covariates. An approach is introduced that combines hypothesis testing based on simultaneous confidence bands with the flexible GAMLSS framework. In particular, GAMLSS is used for the initial parametric model fitting and for the subsequent generation of bootstrap samples. This offers the possibility to account for a plethora of complex multidimensional data structures. The hypothesis is formulated such that rejection of the null hypothesis indicates either the detection of an alert for one covariate at fixed values of the others or the presence of an alert relationship between covariates. The performance of the approach is evaluated in a simulation study that investigates the benefits of the GAMLSS framework. Its practical relevance is demonstrated through an application to data from a study examining the effect of exposure duration on cytotoxicity in primary human hepatocytes.

Authorship:	First author
Coauthors:	Niklas Hagemann and Kathrin Möllenhoff
Contribution statement:	My contributions to this project included developing the method and implementing it in R, including programming, executing the simulation studies and case study, preparing figures and tables, and writing the first draft of the manuscript. The initial idea for the approach was developed by me and Kathrin Möllenhoff. The idea of adjusting the initial model fit and the bootstrap sample generation to a more flexible regression framework was developed by me. Niklas Hagemann advised the use of GAMLSS. Niklas Hagemann and Kathrin Möllenhoff supported the development of the method. All coauthors discussed the contents and structure and revised the manuscript.
Status:	Submitted
Journal:	Preprint available on arXiv
DOI:	arXiv:2604.13264

Estimating effect thresholds and beyond: A flexible framework for multivariate alert detection

L. Ameis^{1,2}, N. Hagemann^{1,2,3}, and K. Möllenhoff^{1,2*}

¹Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne

²Division of Mathematics, Department of Mathematics and Computer Science, University of Cologne

³Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University

Abstract

Evaluating the influence of continuous covariates, like exposure time or dose, on a response variable is a pivotal objective in the assessment of a compound's effect, particularly when determining toxicity in pre-clinical research or pharmacokinetics in clinical trials. The determination of an alert, such as the ED_{50} value, at which a pre-specified threshold of the response variable is crossed, is an important tool for the evaluation process. In practice, response data might be available for combinations of different covariates and the alert depending on both is of interest. In this case, it is crucial to use all available information and extrapolate between cases to ensure the optimal utilization of the data.

In this paper, we introduce a parametric approach that allows alerts to be estimated in a multidimensional setting. For time-dose-response data, for instance, alert doses at a given time can be determined, even when there are no measurements available at that exact time. Likewise, it allows estimation of alert times for a given dose. More generally, the method makes it possible to characterize the complete alert relationship between covariates by leveraging all available data. This is achieved by fitting a parametric model and constructing either a confidence band for the two-dimensional curve given for example a fixed time or dose or by constructing a confidence plane for the three-dimensional model fit. Both are derived by a two-step bootstrap approach and summarized in terms of a hypothesis test that can be adjusted to accommodate a variety of alert types. The initial model fit is achieved by the flexible framework of Generalized Additive Models for Location, Scale and Shape (GAMLSS), which offers the possibility to account for a plethora of complex three-dimensional data structures.

We demonstrate the validity of our approach through a simulation study and present an application to data from a study investigating the relevance of the exposure duration on cytotoxicity in primary human hepatocytes.

1 Introduction

Many pharmacological and toxicological processes are inherently three-dimensional in nature. They depend not only on the administered dose, but also on the time since exposure or, in combination settings, on the joint action of multiple agents. In drug combination studies, this multidimensional relationship can reduce drug resistance and adverse effects while enhancing therapeutic efficacy. This phenomenon has been observed in a variety of research areas, including cancer studies (Mokhtari et al., 2017) and anesthesia research (Lee, 2010).

Similarly, incorporating both exposure time and concentration is essential when assessing compound effects, particularly in the context of determining pharmacokinetics (Lange and Schmidli, 2015) in clinical trials, or analyzing toxicity (Gu et al., 2018) in pre-clinical research.

When it comes to toxicological data, a recent publication suggests the use of parametric approaches (Kappenberg et al., 2023). In particular, the identification of alerts in dose-response or time-response data has benefited from parametric modeling approaches (see Möllenhoff et al. (2022); Kappenberg et al. (2021); Ameis and Möllenhoff (2024)). These methods allow for an interpolation between measured time points or dose levels, thereby improving upon the classical approach of conducting multiple t -tests or the recommended Dunnett test (Hothorn, 2014). With the recent advancements in surface modeling for multidimensional response data, such as in Zhou et al. (2025) and Huusari et al. (2025) for drug-combinations or Duda et al. (2022) and Schüttler et al. (2019) for risk assessment, the extension of alert identification to these settings is a natural methodological next step.

A wide range of alert criteria has been proposed. These include the lowest dose at which a noticeable effect is observed (Jensen et al., 2019) and the lowest measured dose at which the difference between the dose and the control significantly exceeds a critical response threshold (Delignette-Muller et al., 2011). Similarly, one could consider the well-known ED_{50} value, or the more general ED_p ($0 < p < 100$) value (Bretz et al., 2010), an alert corresponding to the dose, where half (or $p\%$) of the maximum predicted effect is observed. For multidimensional mixture effect modeling, Martin-Betancor et al. (2015) introduced a multivariate extension to the Loewe additivity (Loewe, 1953) and, recently, Schürmeyer et al. (2025) constructed efficient designs for identifying effective dose combinations. Both approaches are based on multivariate extensions of the ED_p , thereby extending classical alert definitions for three-dimensional response settings.

Alerts that are identified based on multi-dimensional data incorporate more than one covariate, which is often necessary to obtain a comprehensive understanding of the biological background. For instance, considerable variations in the toxicity of compounds can be observed across varying dose levels and exposure durations in cytotoxicity data (Arbo et al., 2016; Gu et al., 2018; Riddell et al., 1986). The identification of alerts, such as the ED_{50} , exclusively for specific exposure periods, has the potential to result in under- or overestimation of the toxic potential. Furthermore, the identification of a multi-dimensional alert based on multiple exposure durations, in conjunction with varying dose levels, facilitates interpolation and the incorporation of all available data. The latter approach can help avoid overfitting when compared with the fitting of curves to individual exposure durations and the subsequent inference of results (Duda et al., 2023). A second example are drug combinations studies. As previously mentioned, the combination of drugs has been demonstrated to reduce negative and enhance positive effects. The identification of multi-dimensional alerts, such as the multiple effective dose (MED), as for example investigated in (Schürmeyer et al., 2025), allows for the quantification of their combined effect. This, in turn, can assist in the assessment of suitable combinations. For example, known reverse effects of specific dose levels of one of the drugs can be avoided.

In this paper, a model-based approach is proposed that allows the identification of alerts in a three-dimensional data situation, such as, for example, time-dose-response data or the combined effect of two drugs in a drug combination study. Inspired by Möllenhoff et al. (2022) and Kappenberg et al. (2021), this is achieved through the formulation of a hypothesis test, thereby basing alert identification on statistical significance. The test is formulated for either a two-dimensional question or a three-dimensional question that considers a test statistic based on whether the response signal significantly exceeds a pre-specified relevance threshold λ . In the formulation of the hypothesis, the response signal is centered around a reference level, for example, a placebo dose. Depending on the data structure, this might not be necessary and the formulation can be simplified. These simplifications are also presented. In all cases, rejecting the null hypothesis indicates an alert, defined as the smallest dose or time point (per time point or dose level, respectively) for which the null hypothesis is rejected. Depending on the precise formulation of the hypothesis, the test decision is based on the construction of a

lower or upper simultaneous confidence band for the two-dimensional question or a lower or upper simultaneous confidence plane for the three-dimensional question. These are estimated via an extension of the bootstrap approach introduced in Möllenhoff et al. (2022).

The proposed method is specifically designed to accommodate complex three-dimensional data structures. Both the two-dimensional and three-dimensional approaches are based on an initial three-dimensional model fit, for which the framework of Generalized Additive Models for Location, Scale and Shape (GAMLSS) (Rigby and Stasinopoulos, 2005) is utilized. This introduces a high degree of flexibility and allows for a wide range of modeling extensions. In contrast to classical mean-based regression approaches, GAMLSS enable the simultaneous modeling of multiple distributional parameters, including variability. This is particularly relevant when considering heteroscedasticity depending, for example, on the dose level or exposure duration in a complex manner. The GAMLSS approach integrates a multitude of adjustments, such as the inclusion of random effects or spline based modeling, and is continuously developed and well integrated in R (De Bastiani et al., 2026). Therefore, this allows for the generation of bootstrap samples that capture the intricate characteristics of the original data set.

This paper is structured as follows: Initially, the methodology for identifying alerts and alert relationships is introduced based on an estimated simultaneous confidence band or plane, respectively. Subsequently, the approach is evaluated in a simulation study. Finally, the application of the method is demonstrated using a data set from a cytotoxicity study (Gu et al., 2018) that investigates the relevance of the exposure duration in primary human hepatocytes.

2 Model fit

Although the proposed approach is generally applicable to any two or even more continuous variables, we deploy a notation based on its most relevant application of time-dose-response data for notational simplicity. Fundamentally, the idea of this method is to initially fit a parametric regression model to the data, and subsequently evaluate the response by determining an alert of interest based on a hypothesis test. The procedure was inspired by the work of Möllenhoff et al. (2022) and adapted to the more complex nature of a three-dimensional data structure. In this section, the initial model fit is discussed.

Let the data set under consideration consist of n observations in total, collected at $m_t + 1$ different time points, $t_p \in T \subset \mathbb{R}$, where $p = 0, \dots, m_t$, and at $m_d + 1$ different dose levels, $d_q \in D \subset \mathbb{R}$, where $q = 0, \dots, m_d$. Without loss of generality, it is assumed that t_0 and d_0 are the respective reference or placebo levels. Let $n_{p,q}$ denote the number of observations taken at the time-dose combination $(t_p, d_q) \in T \times D$. After a possible preliminary selection step from a pool of candidate models to ensure goodness of fit - using, for example, the AIC (Fahrmeir et al., 2022) as criteria - the initial step for all variations of the proposed approach is to fit a parametric model to the data. There are a multitude of models that adjust for the challenges of complex data structures like autocorrelation or random effects (Box et al., 2015; Gałeczki and Burzykowski, 2012; Demidenko, 2013). In addition to other applications, this is particularly relevant in time-dose-response models, where residual variability can be heteroscedastic and may depend on explanatory variables. Therefore, in this manuscript, the flexible approach offered by Generalized Additive Models for Location, Scale, and Shape (GAMLSS) (Rigby and Stasinopoulos, 2005) is chosen. The GAMLSS framework allows all the parameters of the response variable distribution to be modeled as regression functions of the explanatory variables. In (time-)dose-response models, often normality is assumed (see, e.g., Möllenhoff et al. (2022); Ameis and Möllenhoff (2024)). Hence, we will focus on normally distributed responses, despite the GAMLSS being also applicable to non-normal response variables. Let

$$y_{pq} \sim \mathcal{N}(\mu_{pq}, \sigma_{pq}^2)$$

with $p = 0, \dots, m_t$ and $q = 0, \dots, m_d$. In this case, the GAMLSS framework enables modeling both the mean

(μ) and the standard deviation (σ). Precisely, the regression models under consideration are

$$\begin{aligned}\text{id}(\mu_{pq}) &= \mu_{pq} = f(t_p, d_q, \boldsymbol{\theta}) \\ \log(\sigma_{pq}) &= \eta_{pq} = g(t_p, d_q, \boldsymbol{\vartheta})\end{aligned}\tag{1}$$

which depend on time and dose. Hereby, the identity and the log-function serve as link functions. For an observation y_{pqr} this means

$$y_{pqr} = f(t_p, d_q, \boldsymbol{\theta}) + \varepsilon_{pqr},$$

whereby $\varepsilon_{pqr} \sim \mathcal{N}(0, \sigma_{pq}^2) = \mathcal{N}(0, \exp(g(t_p, d_q, \boldsymbol{\vartheta}))^2)$.

GAMLSS allow for different effect types, including linear, non-linear or spline-based models, and random effects (Stasinopoulos and Rigby, 2008). Here, the emphasis is placed on parametric non-linear regression for the mean (f) and linear regression for the standard deviation (g) for the sake of simplicity. The estimation of GAMLSS is based on maximizing the penalized likelihood as described in, for example, Stasinopoulos et al. (2017). However, as the regression models under consideration, i.e. parametric linear and non-linear models, do not involve penalized terms, the estimation reduces to maximum likelihood estimation (Stasinopoulos and Rigby, 2008).

3 Alert detection

After the initial model fit, the method can be applied in two distinct ways. Firstly, the parametric nature of the approach makes it possible to determine the alert for a fixed dose level or time point, even where no measurements were taken. Secondly, it can be employed to predict the two-dimensional alert relationship between time and dose. In this section, both applications are discussed.

3.1 Hypothesis test: Two-dimensional question

In the following, the approach for a fixed time point $\tilde{t} \in T$ is presented, thereby projecting from three-dimensional data to a two-dimensional question. The procedure can be conducted analogously for a fixed dose level. In this two-dimensional situation, the alert corresponds to the minimum dose, where the response (centered on the response at the placebo level) significantly exceeds a pre-specified relevance threshold λ , with respect to the fixed time point $\tilde{t} \in T$. In terms of a hypothesis test, this can be expressed as follows:

$$\begin{aligned}H_0 : \forall d \in D : |f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})| &\leq \lambda \\ \text{vs. } H_1 : \exists d \in D : |f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})| &> \lambda\end{aligned}\tag{2}$$

for a fixed $\tilde{t} \in T$. Rejecting H_0 means detecting an alert. The test decision is conducted as described in Möllenhoff et al. (2022) by means of the lower simultaneous confidence band to the α -level, thereby soliciting a test at the significance level α . Precisely, let $LB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}})$ denote the lower simultaneous $(1 - \alpha)$ -confidence band of $|f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|$, i.e.

$$\mathbb{P}\{\forall d \in D : LB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}}) \leq |f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|\} \geq 1 - \alpha.\tag{3}$$

If $LB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}}) > \lambda$, H_0 is rejected and the alert dose is defined as the minimum dose, where H_0 is rejected, i.e. $\min\{d \in D : LB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}}) > \lambda\}$. This procedure results in a test at the α -level (Möllenhoff et al., 2022).

Figure 1 (A) depicts a visualization where the black and blue lines represent f and the confidence band, respectively. The solid blue line indicates the confidence band estimated from the two-dimensional approach,

while the dashed line indicates the confidence band inferred from the three-dimensional algorithm described in Section 3.2. The red horizontal line shows the threshold value λ and the red vertical line shows the identified alert.

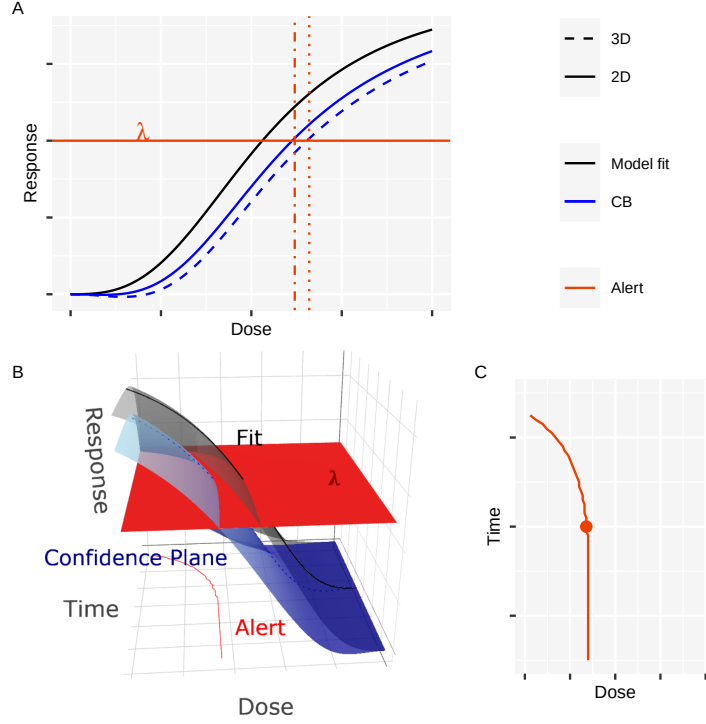


Figure 1: (A) Centered absolute value of the model fit and estimated confidence bands for a fixed time point. The solid blue line indicates the confidence band estimated from the two-dimensional algorithm. The dashed line indicates a confidence band inferred from the three-dimensional analysis. The red horizontal line marks the threshold value. The red vertical lines visualize the estimated alert doses. (B) Visualization of the three-dimensional approach. The centered absolute values of the model fit are indicated by the black plane. The estimated confidence plane is displayed in blue. The threshold λ is indicated by the red hyperplane orthogonal to the *Response*-axis. The time-dose-alert relationship is projected onto the hyperplane spanned by the time-axis and the dose-axis. The black and blue lines exemplify a possible fixed time point. (C) Estimated time-dose-alert relationship resulting from the three-dimensional approach. That is the intersection between the blue surface and the red plain. The dot indicates the alert for the fixed time point.

The centering employed in the formulation of the approach ensures that the expression $|f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|$ is 0 at the reference level, which facilitates a streamlined selection of the relevance threshold λ . Additionally, $|f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|$ increases from 0 at d_0 . Hence, the threshold value is positive and it is tested whether the threshold is exceeded. Defining the alert dose as the minimum dose at which the threshold is exceeded ensures the uniqueness of the identified dose level, even if the increase is non-monotonic.

Depending on the data, the approach can be simplified, when centering to the reference level adds an unnecessary extra step. An illustrative example is cell viability data, as presented in the case study in Section 5. Here, the response values are measured in percentage, start at approximately 100% cell viability and tend to 0%. An alert of interest could be, when the cell viability falls below a value of $p\%$, e.g. 50%, and the relevance threshold would correspondingly be chosen as $p\%$. Centering would result in values of the form $|f(\tilde{t}, d, \boldsymbol{\theta}) - 100|$, which introduces unnecessary complexity into the interpretation. In such cases, the formulation of the hypothesis can be based on $f(\tilde{t}, d, \boldsymbol{\theta})$ instead of $|f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|$. However, it then depends on whether the interest lies in $f(\tilde{t}, d, \boldsymbol{\theta})$ undercutting or exceeding the pre-specified threshold for any $d \in D$. That is

Threshold undercut:

$$H_0 : \forall d \in D : f(\tilde{t}, d, \boldsymbol{\theta}) \geq \lambda \quad \text{vs.} \quad H_1 : \exists d \in D : f(\tilde{t}, d, \boldsymbol{\theta}) < \lambda \quad (4)$$

Threshold exceeded:

$$H_0 : \forall d \in D : f(\tilde{t}, d, \boldsymbol{\theta}) \leq \lambda \quad \text{vs.} \quad H_1 : \exists d \in D : f(\tilde{t}, d, \boldsymbol{\theta}) > \lambda \quad (5)$$

for a fixed $\tilde{t} \in T$. In case (5), the test decision is made using the lower simultaneous confidence band of $f(\tilde{t}, d, \boldsymbol{\theta})$ as described above. In case (4), the test is instead conducted using an upper simultaneous confidence band $UB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}})$ of $f(\tilde{t}, d, \boldsymbol{\theta})$, which is defined analogously to (3) as:

$$\mathbb{P}\{\forall d \in D : UB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}}) \geq f(\tilde{t}, d, \boldsymbol{\theta})\} \geq 1 - \alpha. \quad (6)$$

Analogously to the test decision based on the lower simultaneous confidence band, H_0 is rejected if $UB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}}) < \lambda$ and the alert dose is defined as $\min\{d \in D : UB^\alpha(\tilde{t}, d, \hat{\boldsymbol{\theta}}) < \lambda\}$. As explained above, taking the minimum ensures the unique identification of an alert dose, even if f is non-monotonic.

3.2 Hypothesis test: Three-dimensional approach

The three-dimensional approach is similar overall, but instead of focusing on results for a fixed time point or dose level, it aims to identify the time-dose-alert relationship. Let $(t_0, d_0) \in T \times D$ be the combination of the reference and the placebo level. In this three-dimensional framework, the hypothesis test can be expressed as follows:

$$\begin{aligned} H_0 : \forall (t, d) \in T \times D : |f(t, d, \boldsymbol{\theta}) - f(t_0, d_0, \boldsymbol{\theta})| \leq \lambda \\ \text{vs.} \quad H_1 : \exists (t, d) \in T \times D : |f(t, d, \boldsymbol{\theta}) - f(t_0, d_0, \boldsymbol{\theta})| > \lambda. \end{aligned} \quad (7)$$

Rejecting H_0 again means detecting an alert. Rather than using a confidence band for a two-dimensional curve, the test decision is reached via a lower simultaneous confidence plane $LP^\alpha(t, d, \hat{\boldsymbol{\theta}})$ of $|f(t, d, \boldsymbol{\theta}) - f(t_0, d_0, \boldsymbol{\theta})|$ to the α -level. This confidence plane fulfills

$$\mathbb{P}\{\forall (t, d) \in T \times D : LP^\alpha(t, d, \hat{\boldsymbol{\theta}}) \leq |f(t, d, \boldsymbol{\theta}) - f(t_0, d_0, \boldsymbol{\theta})|\} \geq 1 - \alpha.$$

Conceptually, the idea is the same as above, but defined for the whole plane. Again, H_0 in (7) is rejected if $LP^\alpha(t, d, \hat{\boldsymbol{\theta}}) > \lambda$. The procedure results in a test to the α -level since

$$\begin{aligned} \mathbb{P}_{H_0}\{\exists (t, d) \in T \times D : LP^\alpha(t, d, \hat{\boldsymbol{\theta}}) > \lambda\} &\leq \mathbb{P}_{H_0}\{\exists (t, d) \in T \times D : LP^\alpha(t, d, \hat{\boldsymbol{\theta}}) > |f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|\} \\ &= 1 - \mathbb{P}_{H_0}\{\forall (t, d) \in T \times D : LP^\alpha(t, d, \hat{\boldsymbol{\theta}}) \leq |f(\tilde{t}, d, \boldsymbol{\theta}) - f(\tilde{t}, d_0, \boldsymbol{\theta})|\} \\ &\leq 1 - (1 - \alpha) = \alpha. \end{aligned}$$

A visualization is depicted in Figure 1 (B). Here, the detected alert is not a single dose level indicated by a vertical red line, but rather the time-dose-alert relationship, projected onto the hyperplane spanned by the time-axis and the dose-axis. Additionally, this projection is depicted in Figure 1 (C).

This theory naturally yields the alert for a fixed time point (or dose level) as a special case, which can be determined directly from the estimated confidence plane. The two-dimensional projection of the resulting confidence band is illustrated as the blue dashed line in Figure 1 (A). By nature of construction (see Section 3.3), the resulting confidence band is wider than that estimated using the two-dimensional approach. Simplified hypotheses are analogous to (4) and (5).

3.3 Testing procedure

In principle, the test can be conducted using any method of constructing a lower or upper simultaneous confidence band for a two-dimensional curve, or a lower or upper simultaneous confidence plane in three dimensions, respectively. This manuscript utilizes and extends the procedure introduced in Möllenhoff et al. (2022) due to its natural flexibility. Let $f(t, d, \hat{\theta})$ be the estimated parametric model fit of the mean. Let

$$\Delta(t, d, \hat{\theta}) = \begin{cases} |f(\tilde{t}, d, \hat{\theta}) - f(\tilde{t}, d_0, \hat{\theta})|, & \text{two-dimensional case, centered} \\ f(\tilde{t}, d, \hat{\theta}), & \text{two-dimensional case, simplified} \\ |f(t, d, \hat{\theta}) - f(t_0, d_0, \hat{\theta})|, & \text{three-dimensional case, centered} \\ f(t, d, \hat{\theta}), & \text{three-dimensional case, simplified} \end{cases}$$

depending on whether the two- or three-dimensional hypothesis is regarded and whether it is centered or simplified. The lower simultaneous confidence plane is constructed based on an estimator for the standard deviation of $\Delta(t, d, \hat{\theta})$ and a quantile c , such that

$$\mathbb{P} \left\{ \max_{(t,d) \in T \times D} \frac{\Delta(t, d, \hat{\theta}) - \Delta(t, d, \theta)}{\hat{\sigma}_{\Delta(t,d,\hat{\theta})}} \leq c_{lower, 3dim} \right\} = 1 - \alpha$$

in the three-dimensional case. For the two-dimensional case, the maximum is taken over the subset $\{\tilde{t}\} \times D$. Hence, the lower simultaneous confidence band or plane can be constructed as

$$L^\alpha(t, d, \hat{\theta}) := \Delta(t, d, \hat{\theta}) - c_{lower, 2dim/3dim} \hat{\sigma}_{\Delta(t,d,\hat{\theta})}, \quad (8)$$

whereby L denotes either LB or LP . It holds $LB^\alpha(\tilde{t}, d, \hat{\theta}) \geq LP^\alpha(\tilde{t}, d, \hat{\theta})$, since $c_{lower, 2dim} \leq c_{lower, 3dim}$, as the maximum difference between the true and estimated value is taken over a subset in the two-dimensional case. Naturally, this can also be explained because the curve for a fixed time point $\tilde{t}$ is part of the three-dimensional plane, and a simultaneous confidence plane to the α -level also encapsulates this curve to the α -level. However, if only a fixed time point is of interest, the approach based on the confidence plane will inherently yield a more conservative test.

In order to estimate $LB^\alpha(\tilde{t}, d, \hat{\theta})$ or $LP^\alpha(t, d, \hat{\theta})$, it is necessary to estimate $c_{lower, 3dim/2dim}$ and $\hat{\sigma}_{\Delta(t,d,\hat{\theta})}$. This is achieved via a nested two-level bootstrap procedure as introduced in Möllenhoff et al. (2022). This approach resembles the classic bootstrap-t intervals introduced in Efron and Tibshirani (1994), but is applied to the entire plane or curve. Calculating the maximum for the estimation of c ensures that $LB^\alpha(\tilde{t}, d, \hat{\theta})$ or $LP^\alpha(t, d, \hat{\theta})$ are blown up from enclosing the curve or plane pointwise to simultaneously enclosing it at α -level. Precisely, B_1 bootstrap samples are generated similar to the original data set (for details see Section 3.4). The empirical standard deviation $\hat{\sigma}_{\Delta(t,d,\hat{\theta})}$ is then estimated from $\Delta(t, d, \hat{\theta}_1^*), \dots, \Delta(t, d, \hat{\theta}_{B_1}^*)$. Additionally, for each generated first-level bootstrap sample $l = 1, \dots, B_1$, a second level of B_2 bootstrap samples is generated. The sample generation step is analogous, but the first-level bootstrap sample is used instead of the original data set as a basis. From each set of B_2 second-level bootstrap samples, the empirical standard deviation $\hat{\sigma}_{\Delta(t,d,\hat{\theta}_l^*)}$ is estimated. These are then used to calculate

$$D^{*,l} = \max_{(t,d) \in T \times D} \frac{\Delta(t, d, \hat{\theta}_l^*) - \Delta(t, d, \hat{\theta})}{\hat{\sigma}_{\Delta(t,d,\hat{\theta}_l^*)}},$$

from which c is estimated as the empirical $(1 - \alpha)$ -quantile.

Sample generation is identical for the two- and three-dimensional approaches. The other steps are similar, with the only difference being that $t = \tilde{t}$ is fixed for the two-dimensional case. The upper simultaneous confidence band or plane, $UB^\alpha(\tilde{t}, d, \hat{\theta})$ or $UP^\alpha(t, d, \hat{\theta})$, is constructed and estimated analogously.

3.4 Generation of bootstrap samples and faster algorithm

There is a fundamental interconnection between the model assumption, estimation, and bootstrap sample generation. As discussed in Section 2, fitting GAMLSS to the data results in estimators $\hat{\theta}$ and $\hat{\vartheta}$ for the regression models of the mean μ and standard deviation σ , respectively. Accordingly, an observation y_{pqr}^* at $(t, d) \in T \times D$ for a first-level bootstrap sample is generated by drawing errors $\varepsilon_{pqr}^* \sim \mathcal{N}(0, \exp(g(t, d, \hat{\vartheta})^2))$ and calculating

$$y_{pqr}^* = f(t, d, \hat{\theta}) + \varepsilon_{pqr}^*.$$

Let $\hat{\theta}_l^*$ and $\hat{\vartheta}_l^*$, $l = 1, \dots, B_1$, denote the estimated parameters from a first-level bootstrap sample. An observation of a second-level bootstrap is generated similarly by drawing errors $\varepsilon_{pqr}^{**} \sim \mathcal{N}(0, \exp(g(t, d, \hat{\vartheta}_l^*)^2))$ and calculating

$$y_{pqr}^{**} = f(t, d, \hat{\theta}_l^*) + \varepsilon_{pqr}^{**}.$$

The estimation procedure for GAMLSS requires more computational effort than a simple OLS or maximum likelihood (ML) estimation. Therefore, reducing the computation time might be of interest if the approach is applied to a large number of data sets. Due to the model assumptions, both $\hat{\theta}^{**}$ and $\hat{\vartheta}^{**}$ are estimated for each of the second-level bootstrap samples, respectively. However, only $\hat{\theta}^{**}$ is utilized further in the algorithm. Therefore, it is possible to simplify the estimation in this step and use OLS/ML to estimate only $\hat{\theta}^{**}$ assuming iid. error terms $\varepsilon_{pqr} \sim N(0, \sigma^2)$. This more simplistic assumption has the potential to compromise the precision of the model fit, yet reduces the time required for computation.

In general, both algorithms produce similar results. Due to the sake of brevity, the results obtained by applying the fast algorithm are presented in this manuscript. The results of the normal algorithm are presented in the supplementary material. Note that both algorithms are identical if a constant standard deviation is assumed.

4 Simulation Study

4.1 Setting and data generation

In order to evaluate the proposed approach, a simulation study is conducted. To get a broad view of the method's performance, two distinct scenarios are considered, each based on a different data type. Scenario 1 is based on the time-dose-response study analyzing cytotoxicity presented as a case study in Section 5. The response variable is cell viability, measured as a percentage, ranging from around 100% to 0%. Measurements were taken at 3 time points: day $t_0 = 1$, day $t_1 = 2$, and day $t_2 = 7$. For each time point, 6 equidistant dose levels on a log scale with base $= \sqrt{10}$ were measured. The alert of interest is 50% cell viability. Therefore, the simplified assumption for undercutting a threshold as presented in (4) applies and an upper simultaneous confidence band/plane is estimated.

For the parametric model, the *td2pLL* model introduced in Duda et al. (2022) is selected. See the supplementary material for details on the parametrization. To evaluate the benefits of the GAMLSS framework, four different cases are compared: constant standard deviation (simple) and a standard deviation represented by the linear model

$$\log(\sigma) = \beta_{Intercept} + \beta_{Dose}Dose + \beta_{Dose^2}Dose^2 + \beta_{Time}Time + \beta_{Dose^2 \cdot Time}Dose^2 \cdot Time \quad (9)$$

with small, medium and large coefficients. The formulation of the linear model is motivated by the findings of Duda et al. (2022). The parameters used for the simulation are obtained from Case 2, as presented in Section 5. Two sample sizes, $n = 250$ and $n = 54$, are considered.

All cases are analyzed applying two distinct assumptions: one assuming a constant standard deviation, and the other assuming the more complex linear dependency shown above. Both the three-dimensional and two-dimensional approaches are utilized. For the two-dimensional approach, the fixed time point of $t = 4$ is selected. This choice is consistent with the fixed time point selected in the case study in Section 5.

Scenario 2 is based on the simulation study presented in Schürmeyer et al. (2025), where the optimal design for identifying effective dose combinations within the context of drug combination studies was investigated. The example illustrates a clear positive interaction between two drugs, here denoted Dose 1 and Dose 2 for simplicity, and an overall increase in response. Precisely, this means that the response of the individual drugs is increasing, and that the combination of these drugs produces a response that exceeds the maximum response of the individual drugs. The underlying parametric model is a drug-combination model, whereby both compounds follow the form of an *E_{max}* model (Bornkamp et al., 2009) (for details see Schürmeyer et al. (2025) or the supplementary material). The alerts of interest are $p\%$ multivariate effective dose (MED) contour lines defined as

$$\text{MED}_p(\boldsymbol{\theta}) = \left\{ (d_1, d_2) \in D_1 \times D_2 : \frac{f(d_1, d_2, \boldsymbol{\theta}) - \min_{(d_{1,0}, d_{2,0}) \in D_1 \times D_2} f(d_{1,0}, d_{2,0}, \boldsymbol{\theta})}{R_{\max}} = \frac{p}{100} \right\}$$

with $R_{\max} = \max_{(d_1, d_2) \in D_1 \times D_2} f(d_1, d_2, \boldsymbol{\theta}) - \min_{(d_1, d_2) \in D_1 \times D_2} f(d_1, d_2, \boldsymbol{\theta})$. This concept is a two-dimensional extension of the *ED_p* value, whereby the predicted effect is calculated from the minimum and maximum over all dose pairs instead of from the lower and upper asymptotes. Since the alerts can be interpreted without centreing, the simplified assumption described in (5) can be applied and a lower simultaneous confidence band/plane is fitted. The scenario is analyzed using both the three-dimensional and two-dimensional approaches.

To evaluate the influence of the design, two different designs as presented in Schürmeyer et al. (2025) are regarded. Precisely, the factorial 3×3 design and the D-optimal design for (80, 90)% MED contour lines as identified in the manuscript are selected. Accordingly, the 80% MED contour line is chosen as the alert of interest. The exact dose levels for both drugs for which measurements were taken depend on the design. However, in both cases the dose levels range from $d_{1,0} = 0$ to $d_{1,m_{D_1}} = 10$ for Dose 1 and from $d_{2,0} = 0$ to $d_{2,m_{D_2}} = 12$ for Dose 2. For the two-dimensional approach, we choose to fix the dose level of Dose 1 at $\tilde{d}_1 = 7$ for the demonstration, since no measurements were taken for this dose in either design scheme. The sample sizes $n = 152$, $n = 90$ and $n = 45$ are considered. Both scenarios are depicted in Figure 2. For details on the precise parameters, see Table 1.

The simulated datasets are generated as described in Section 3.4. For each combination of scenario, specification, and number of observations, a total of 1,000 simulation runs are conducted. For the proposed approach, a large value ($B_1 = 500$) is chosen for the estimation of the quantiles in the first-level bootstrap, and a small value ($B_2 = 25$) is chosen for the estimation of the standard error in the second-level bootstrap. Thus, the number of generated bootstrap samples overall is $500 \times 25 = 12500$. As recommended in Efron and Tibshirani (1994) and utilized satisfactorily in Möllenhoff et al. (2022) and Ameis and Möllenhoff (2024), these choices are large enough to ensure precise estimations, yet small enough to avoid excessive computational effort.

Outcomes of the two-dimensional approach are the number of rejections of H_0 , which corresponds to the number of detected alerts, and the median and standard deviation of the identified alert dose (of Dose 2). For the three-dimensional approach, one outcome is also the number of rejections of H_0 . Additionally, the estimated time-dose-alert relationships is examined. Specifically, two main characteristics are examined: First, the accuracy of the time period or dose levels of Dose 1 for which alerts were identified. Second, the accuracy of the identified alerts.

Let $\tilde{T}_{true}$ denote the true time period or span of dose levels of Dose 1 for which alerts exist. In Scenario 1, $\tilde{T}_{true} = [1.1, 7]$. The true response does not fall below the threshold value only at $t = 1$. In Scenario 2, $\tilde{T}_{true} = [3.6, 10]$. The true response does not exceed the threshold value for dose levels of Dose 1 in the interval $[0, 3.5]$. However, the time period or span of dose levels of Dose 1 estimated from the simulation runs, denoted

Table 1: Parameters of the simulation scenarios. For details on the design see Table A in the supplementary material. Note that a link function is applied for the standard deviation as described in Section 3.4. For the model parameterizations we refer to Section 1 in the supplementary material.

Scenario	Parameters θ	Parameters ϑ	Observations n	Design
1 - Full - Simple	2.9,5.9,1.8,4.8	2.081	250	Full
1 - Full - Small	2.9,5.9,1.8,4.8	2.081,0.162,-0.019,0.06,-0.001	250	Full
1 - Full - Medium	2.9,5.9,1.8,4.8	2.081,0.215,-0.025,0.08,-0.002	250	Full
1 - Full - Large	2.9,5.9,1.8,4.8	2.081,0.269,-0.032,0.1,-0.002	250	Full
1 - Reduced - Simple	2.9,5.9,1.8,4.8	2.081	54	Reduced
1 - Reduced - Small	2.9,5.9,1.8,4.8	2.081,0.162,-0.019,0.06,-0.001	54	Reduced
1 - Reduced - Medium	2.9,5.9,1.8,4.8	2.081,0.215,-0.025,0.08,-0.002	54	Reduced
1 - Reduced - Large	2.9,5.9,1.8,4.8	2.081,0.269,-0.032,0.1,-0.002	54	Reduced
2 - Factorial 3x3 - N152	0,80,3,120,10,0.02	3.401	152	Factorial 3x3
2 - Factorial 3x3 - N90	0,80,3,120,10,0.02	3.401	90	Factorial 3x3
2 - Factorial 3x3 - N45	0,80,3,120,10,0.02	3.401	45	Factorial 3x3
2 - D-optimal - N152	0,80,3,120,10,0.02	3.401	152	D-optimal
2 - D-optimal - N90	0,80,3,120,10,0.02	3.401	90	D-optimal
2 - D-optimal - N45	0,80,3,120,10,0.02	3.401	45	D-optimal

$\tilde{T}_{est}$, can differ from these intervals. The accuracy of the estimation is measured in recall, precision, onset error and offset error. Precisely:

$$\begin{aligned}
\text{Recall}_{est} &= \frac{|\tilde{T}_{true} \cap \tilde{T}_{est}|}{|\tilde{T}_{true}|} \\
\text{Precision}_{est} &= \frac{|\tilde{T}_{true} \cap \tilde{T}_{est}|}{|\tilde{T}_{est}|} \\
\text{Onset error}_{est} &= \min(\tilde{T}_{est}) - \min(\tilde{T}_{true}) \\
\text{Offset error}_{est} &= \max(\tilde{T}_{est}) - \max(\tilde{T}_{true})
\end{aligned}$$

Recall measures how much of the truth the estimates cover, while precision measures how correct the estimates are (Fawcett, 2006; Manning, 2008). Hereby, $\tilde{T}_{true}$ and $\tilde{T}_{est}$ are discretized by superimposed grid and both values are calculated from the respective cardinalities. These values are examined primarily to identify potential bulks in the estimated confidence planes that could result in the estimated $\tilde{T}_{est}$ not being connected. The onset error and offset error measure the accuracy of the earliest and latest estimated time points or smallest and largest estimated dose levels of Dose 1 for which alerts are detected, respectively.

The accuracy of the estimated alerts is evaluated by means of the root mean squared error (RMSE) for the overlap $\tilde{T}_{true} \cap \tilde{T}_{est}$. Precisely, let $d_{est}(t)$ be the identified alert for $t \in \tilde{T}_{true} \cap \tilde{T}_{est}$ and $d_{true}(t)$ the truth. Then:

$$\text{RMSE}_{est} = \sqrt{\frac{1}{|\tilde{T}_{true} \cap \tilde{T}_{est}|} \sum_{t \in \tilde{T}_{true} \cap \tilde{T}_{est}} (d_{est}(t) - d_{true}(t))^2}$$

It is important to note that in the context of a simulation run, the initial model fit may not converge. This was observed on a single occasion for Scenario 1 with the reduced design and medium simulated standard deviation. The run was excluded from further analysis. All simulation scenarios were run on an HPC-cluster employing R version 4.4.1.

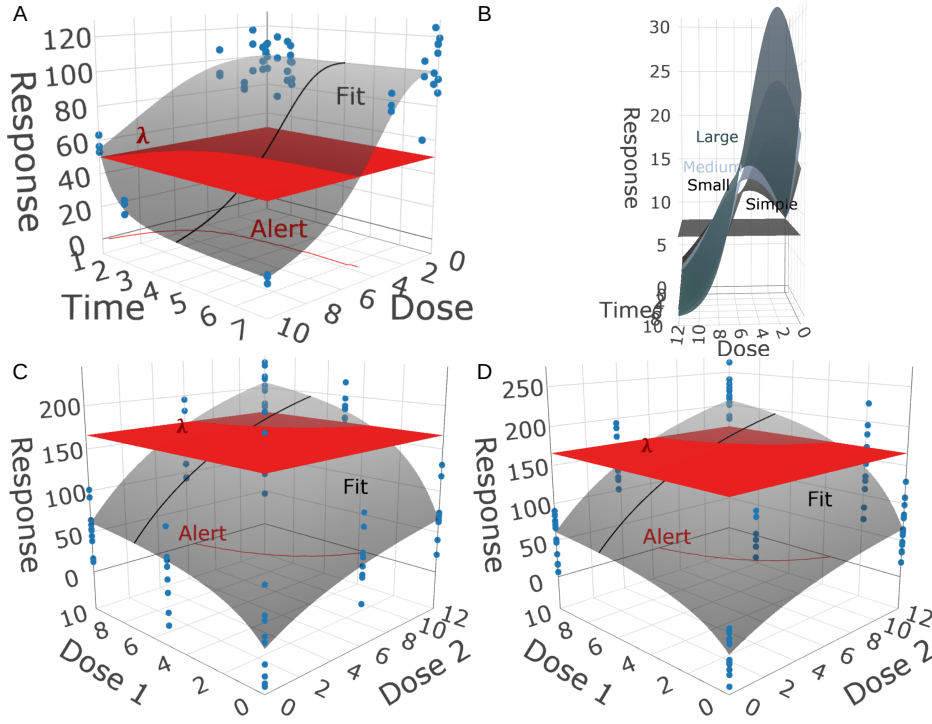


Figure 2: (A) True mean response of Scenario 1. (B) Comparison of the simple, small, medium and large linear models used to generate the standard deviation. (C) Scenario 2 with the factorial 3×3 design. (D) Scenario 2 with the D-optimal design. The black plane in (A), (B) and (C) indicate the true response planes. The black lines indicate the two-dimensional curve for the fixed time point $\tilde{t} = 4$ or dose level of Dose 1 $\tilde{d}_1 = 7$, respectively. The red plane displays the respective thresholds. The intersection between the threshold plane and the true response marks the true alert relationship between time and dose or dose and dose. For better visibility, this relationship is projected to the 0-response axis. The blue dots indicate an example of simulated measurements to demonstrate the desing.

4.2 Results of the two-dimensional hypothesis

In this section, the results of the alerts at the fixed time point of $\tilde{t} = 4$ for Scenario 1, and at the fixed dose level of $\tilde{d}_1 = 7$ for Scenario 2, as estimated by the two-dimensional confidence band or the three-dimensional confidence plane, are presented. For Scenario 1, only the results of the fast algorithm are discussed here. A comparison to the normal algorithm is provided in the supplementary material. Overall, the results are similar. There is no such differentiation for Scenario 2.

Figure 3 and Figure A in the supplementary material, as well as Tables B-E in the supplementary material show the proportion of rejections of H_0 in 1,000 runs, which corresponds to the proportion of detected alerts. As expected, the number of rejections by the two-dimensional approach is greater than or equal to that by the three-dimensional approach, since $c_{3dim} \geq c_{2dim}$. For Scenario 1 with the full design assuming a constant standard deviation, H_0 is always rejected. When the complex structure of the standard deviation is assumed correctly, H_0 is also always rejected, whereas misspecification in the simple case leads to rejection in at least 99.8% runs. For the reduced design, the complex assumption results in rejections in at least 99.9% runs. For the simple and small cases, H_0 is always rejected when assuming a constant standard deviation. For the medium and large cases, however, the number of rejections is slightly smaller, but never dropping below 97.8%.

For Scenario 2 with the factorial 3×3 design a clear trend of fewer rejections is observed as the total number of observations decreases, dropping to 63.9% for the two-dimensional approach and 38.6% for the three-dimensional

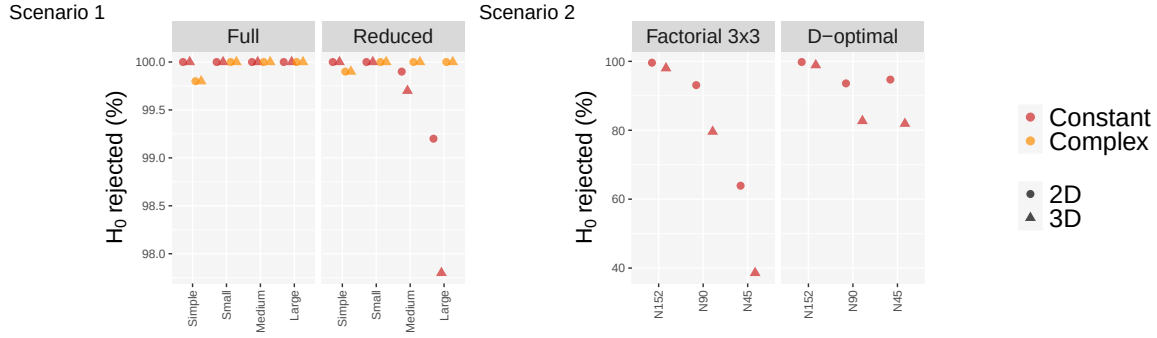


Figure 3: Proportion of rejections of H_0 in both scenarios.

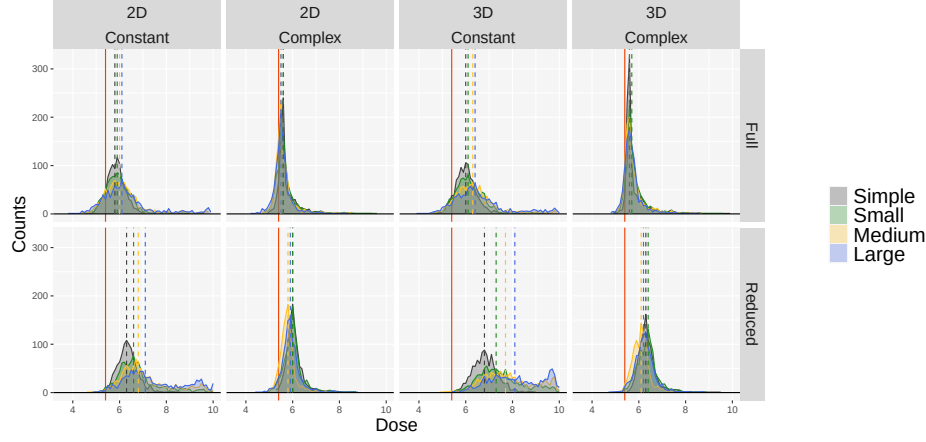
approach. Interestingly, this trend is not observed for the D-optimal design. Although the number of rejections is higher for $n = 152$ observations, it remains almost constant for $n = 90$ and $n = 45$ observations. Additionally, the overall number of rejections is higher for the D-optimal design, exceeding at least 93.6% for the two-dimensional approach and 81.9% for the three-dimensional approach.

Figures 4 (A) and 5 (A) depict the number of the identified alerts per dose level (of Dose 2) and their respective medians for the different cases, compared to the true alert. The medians and the truth are indicated by vertical dashed lines and a solid line, respectively. The colors represent the simulated scenarios for the full (upper row) or reduced (lower row) design. The blocks show the results for either the two-dimensional or three-dimensional approach, assuming a constant or complex structure for the standard deviation. For Scenario 1, it is evident that the medians are closest to the truth for the full design with the two-dimensional approach, as expected. Furthermore, a decline in precision is observed as the standard deviation increases. These trends are visibly less pronounced for the complex assumption. Even in the simple case, the medians are closer to the truth. Additionally, the count distribution is visibly narrower. This clearly shows the benefits of the GAMLSS framework.

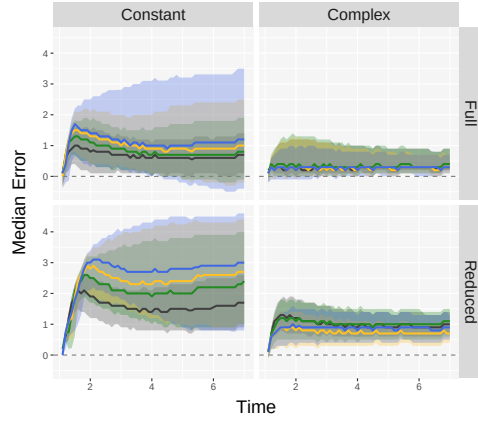
A similar trend is observed for the median and the number of observations in Scenario 2 with the factorial 3×3 design. However, the results for $n = 90$ and $n = 45$ are very similar for the D-optimal design. Additionally, the distribution appears slightly narrower in these cases. A general observation of the results for Scenario 2 reveals the presence of a bias, that is not as evident in the results of Scenario 1. As discussed in Section 2, the assumption of a constant standard deviation might not properly reflect the complex structure of three-dimensional data and lead to a larger variation in simulated data sets. This effect occurs at two levels: Firstly, on the level of the simulated data sets themselves. Secondly, on both levels of the bootstrap procedure. Consequently, variability is amplified throughout the entire simulation setting. This likely impacts on the overall results and reinforces the method's tendency to underestimate the alert already observed for the simpler two-dimensional case in Möllenhoff et al. (2022). This further supports the recommendation outlined in Section 2 to use GAMLSS for the modeling, in order to reduce the technical variability in favor of the inclusion of variability found in the data itself. However, this was not possible to include in Scenario 2 since the parameters were extracted from an external source without a real data basis. To properly evaluate the procedure's performance and to do justice to the complex nature of three-dimensional data an additional simulation setting similar to the complex case in Scenario 1 for Scenario 2 would be necessary. These findings are supported by the observation of a similar, albeit less pronounced, effect in the comparable sub-scenario of Scenario 1. This sub-scenario is the simple case in which the standard deviation is correctly assumed to be constant. The results for median and standard deviation are deferred to Tables B-E in the supplementary material.

Scenario 1 – Results

A



B



C

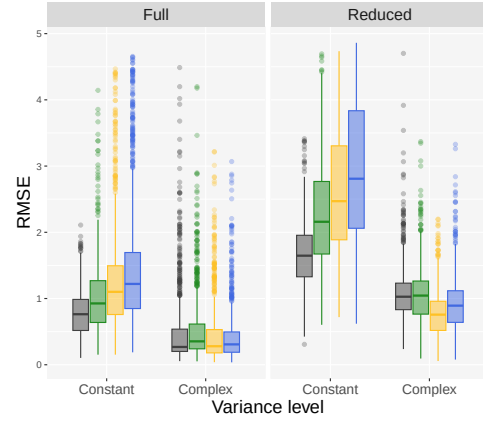


Figure 4: (A) Counts of identified alerts at different dose levels for the two-dimensional approach. Each block contains the results of a model assumption regarding σ (constant or complex) for the full or reduced design, respectively. The colors differentiate between the various simulation scenarios with simple, small, medium and large standard deviation. The red vertical line indicates the true alert. The dashed vertical lines show the respective medians. (B) Median error of the identified alert doses along the time-axis for the three-dimensional approach. Each block contains the results of a model assumption regarding σ (constant or complex) for the full or reduced design, respectively. The solid line indicates the median. The areas indicate the 10% to 90% quantiles. (C) Boxplots of the RMSEs for the three-dimensional approach.

4.3 Results of the three-dimensional hypothesis

In this section, the results for the three-dimensional approach are discussed. Tables F-I in the supplementary material show the proportions of rejections of H_0 in 1,000 simulation runs as well as the mean and standard deviation of the recall, precision, onset error and offset error. Similar trends in the number of rejections of H_0 are observed for three-dimensional hypotheses as for two-dimensional hypotheses. The mean recall for Scenario 1 is consistently high, with values exceeding 94.9% for the full design and 86.3% for the reduced design. In all the cases that were considered, and for all the different assumptions, the standard deviation of the recall is observed to exceed 5% once. The precision is even higher, almost 100% with a standard deviation of 0.1% or smaller. It can be observed that the mean offset error is consistently smaller than the onset error. In fact, it is almost always 0. This is possibly due to the offset equaling the largest time point observed in the study.

Scenario 2 – Results

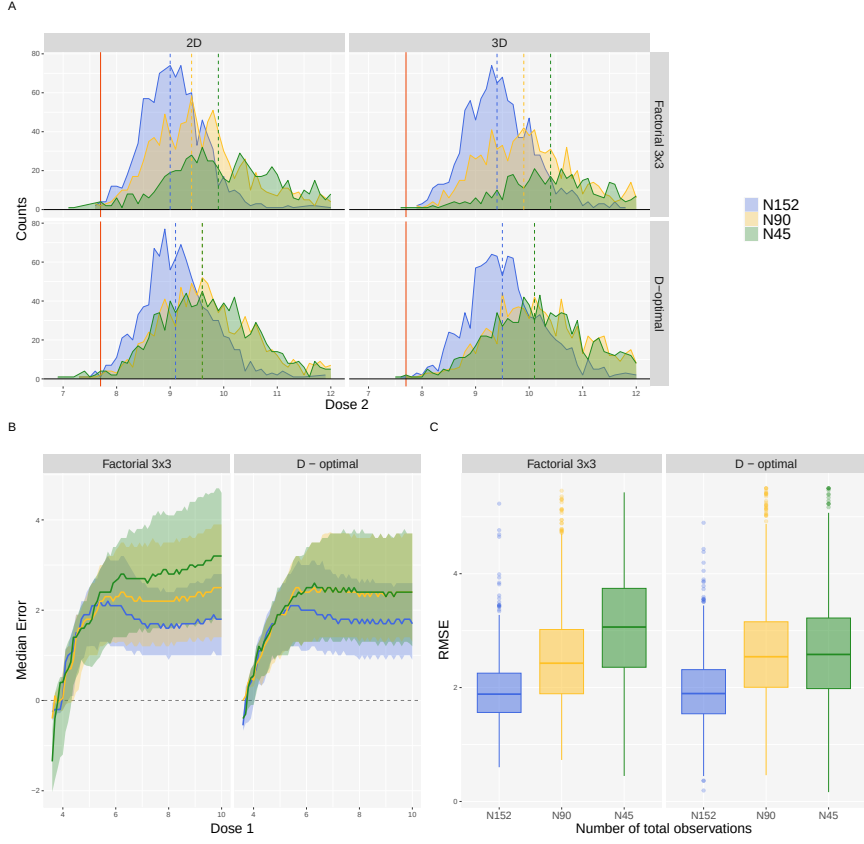


Figure 5: (A) Counts of identified alerts at different levels of Dose 1 for the two-dimensional approach. Each block contains the results of the three different total numbers of observations. The red vertical line indicates the true alert. The dashed vertical lines show the respective medians. (B) Median error of the identified alerts of Dose 2 along the *Dose 1*-axis for the three-dimensional approach. Each block contains the results of the three different total numbers of observations. The solid line indicates the median. The areas indicate the 10% to 90% quantiles. (C) Boxplots of the RMSEs for the three-dimensional approach.

Assuming a complex structure for the standard deviation results in a smaller mean onset error than assuming a constant standard deviation, except in a single case (Full, Simple, Normal), and the error is subject to smaller fluctuations between the cases overall. For the constant assumption, an increasing trend with larger simulated standard deviations can be observed. Both again show the benefit of the utilization of the GAMLSS framework. However, the standard deviation of the onset error is slightly larger overall for the complex assumption in the full design. This is not the case for the reduced design, where the standard deviation for the complex assumption is smaller, except in the simple case. For Scenario 2, smaller means for recall and onset error are observed with the D-optimal design than with the factorial 3×3 design, while the respective standard deviations are similar in size. The results for precision and offset error are comparable to those for Scenario 1.

Figure 4 (B) and Figure 5 (B) depict the median error for the estimated alert dose (of Dose 2, respectively) and the 10% and 90% quantiles per time point (dose level of Dose 1, respectively). For Scenario 1, it is evident that the median error is smaller, and the quantiles are more proximate under the complex assumption. In Scenario 2, the discrepancy between the designs is less pronounced. However, the median error for the D-optimal design appears to be less significant for higher doses of Dose 1 in comparison to the factorial 3×3 design. The results for Scenario 2 again show the bias discussed above for the two-dimensional results.

Figure 4 (C) and Figure 5 (C) depict boxplots of the RMSE for the various cases. For Scenario 1, the boxplots indicate smaller values overall for the complex assumption. The size of the boxplots for the constant assumption increases with the simulated standard deviation, which, in this case, can be interpreted as a degree of misspecification, since the coefficients in (9) except for the intercept differ more from 0. The size of the boxplots remains constant when assuming a complex structure, thereby underscoring the benefit of the GAMLSS framework. For Scenario 2, the absence of a trend is again observed in the results of D-optimal designs for the two smaller numbers of observations, which is clearly visible for the factorial 3×3 design.

5 Case Study

In this section, an application of the proposed approach is presented. For illustration, the data set generated in Gu et al. (2018) is used, in which a total of 30 compounds were assessed in order to ascertain the relevance of the exposure duration on cytotoxicity in primary human hepatocytes. Precisely, the goal was to determine whether additional information can be gained from a prolonged incubation period beyond the usually used 1 to 2 days (Arbo et al., 2016; Ghallab et al., 2016). As demonstrated by Riddell et al. (1986), substantial differences in the toxicity of compounds can be observed between 1 and 3 days. In the study of Gu et al. (2018), an additional third exposure period of 7 days was chosen. Based on statistical significance, the proposed approach can assist in determining whether a shorter time period would have sufficed in some cases. The following analysis focuses on the results generated for *Aspirin* (*ASP*).

For *ASP*, cell viability data is available for 6 concentrations and exposure durations of 1, 2 and 7 days. The concentrations were selected equidistant on a log scale with base $= \sqrt{10}$. The experimental design comprises 3 biological replicates (*Donors*) and 4 or 8 measurements per combination of donor, exposure duration, and concentration. The data is normalized as described in Duda et al. (2022). In addition to the full data set, this section includes an analysis of an artificially reduced data set. This is used to demonstrate the approach in a setting with a smaller number of observations. The reduction is achieved by calculating the mean response of each donor for all combinations of exposure duration and concentration, and treating it as a single measurement. Note that in the following, the terms *time* and *dose* are used instead of *exposure duration* and *concentration* to ensure consistency with the rest of this manuscript.

For cell viability, it can be assumed that following normalization, the response is 100% at the control and tends to 0% at higher concentrations. Hence, the simplified approach for the monotonically decreasing case as described in (4) can be applied. As a threshold value, 50% cell viability was chosen. In this case, the corresponding alert is the value ED_{50} , which is frequently selected to evaluate the potency of a compound (Sebaugh, 2011). For the two-dimensional analysis, the fixed time point was set at 4 days, which is one day longer than the longest time period examined in Riddell et al. (1986) and recommended for many tests in ICCVAM (2001). For the parametric model fitting of the mean, the models proposed in Duda et al. (2022) (*td2pLL*) and Schüttler et al. (2019) (denoted *schuettler* in the following) are selected as competitors, see the supplementary material for details on the parametrization. The following two linear models for the standard deviation are considered:

$$\text{Constant: } \log(\sigma) = \beta_{\text{Intercept}}$$

$$\text{Complex: } \log(\sigma) = \beta_{\text{Intercept}} + \beta_{\text{Dose}} \text{Dose} + \beta_{\text{Dose}^2} \text{Dose}^2 + \beta_{\text{Time}} \text{Time} + \beta_{\text{Dose}^2 \cdot \text{Time}} \text{Dose}^2 \cdot \text{Time}$$

Preliminary investigations indicated that the donor’s influence is negligible and, consequently, has been excluded from consideration. Note that the formula of the complex assumption was used in Duda et al. (2022) to generate a simulation study based on Gu et al. (2018). Also, note that the link function is applied in both cases as described in Section 3.4. By comparing the simple assumption with the more complex alternative, the benefits of the flexibility introduced by the proposed method can be evaluated. Figures D-I in the supplementary material depict the estimated linear models. Model selection was conducted using the AIC (Fahrmeir et al., 2022).

However, given the small differences in AICs, the findings obtained from fitting the respective other parametric model are shown in the supplementary material for comparison. Overall, these results are very similar to those presented in the following.

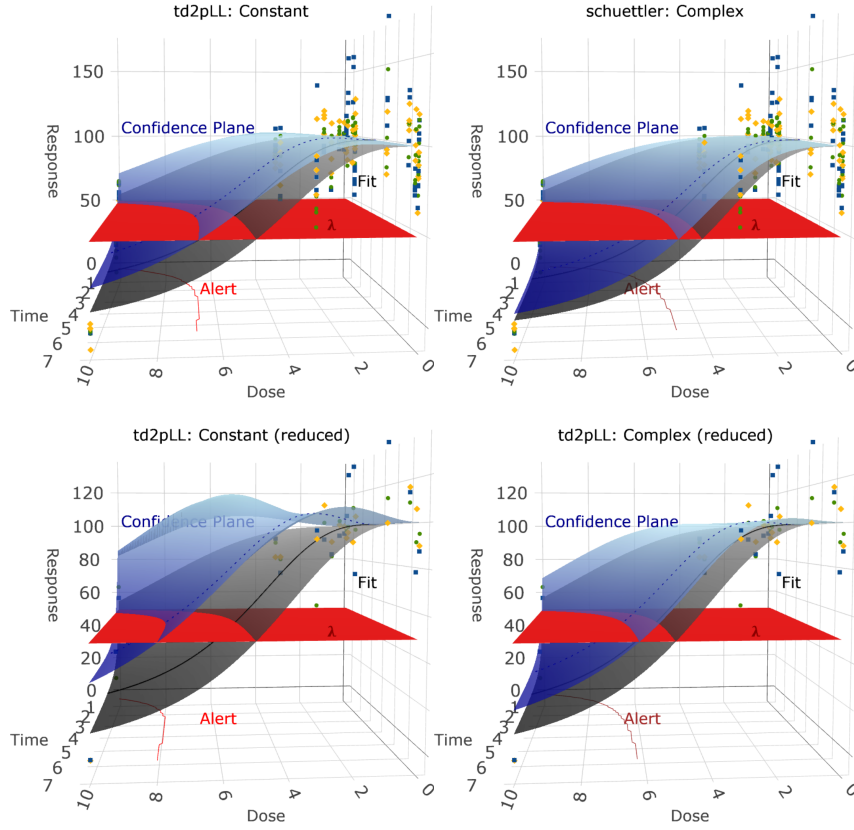


Figure 6: Visualization of the results obtained from the three-dimensional approach. The upper row displays the results for the full data set, the lower row the results for the artificially reduced data set. The dots represent the measured data points, which have been colored according to the donor. The black plane indicates the model fit. The blue plane represents the estimated confidence plane. The fixed time point $\tilde{t} = 4$ is highlighted in both planes. The alert threshold λ is displayed as a red hyperplane that is orthogonal to the *Response*-axis. The estimated alert relationship is projected onto the plane spanned by the dose-axis and the time-axis intersecting the *Response*-axis at the origin.

Figure 6 displays the model fit and the estimated confidence plane for each case. The fixed time point $\tilde{t} = 4$ is highlighted. The red hyperplane indicates the chosen threshold value of 50%. The significant time-dose-alert relationship identified is projected onto the hyperplane spanned by the dose-axis and the time-axis. Overall, it can be observed that the estimated confidence plane derived from the more complex model encloses the model fit more smoothly and identifies smaller alert doses for earlier time points. The former effect is especially visible for the reduced data set. The estimated confidence plane assuming the constant model demonstrates a discernible bulk for shorter time periods, which is not evident in the confidence plane estimated from the complex model. On the left, Figure 7 displays the two-dimensional results for $\tilde{t} = 4$. As previously observed by means of the definition, the estimated confidence band projected from the three-dimensional approach is wider and identifies larger alert doses than the confidence band derived from the two-dimensional approach. However, the difference is small. This indicates only a minor loss in accuracy of the three-dimensional approach on the two-dimensional level.

For the full data set, the identified alerts range from 6.4 (2D) to 6.8 (3D) for the simple model and from 5.2 (2D) to 5.4 (3D) for the complex model. The alerts identified for the artificially reduced data set are larger and extend over a wider interval, as was expected due to the uncertainty introduced by a smaller number of observations. For the constant model, the range is from 6.9 (2D) to 7.9 (3D), whereas for the complex model it is from 6.2 (2D) to 6.7 (3D). All identified alerts are collected in Tables J-K in the supplementary material.

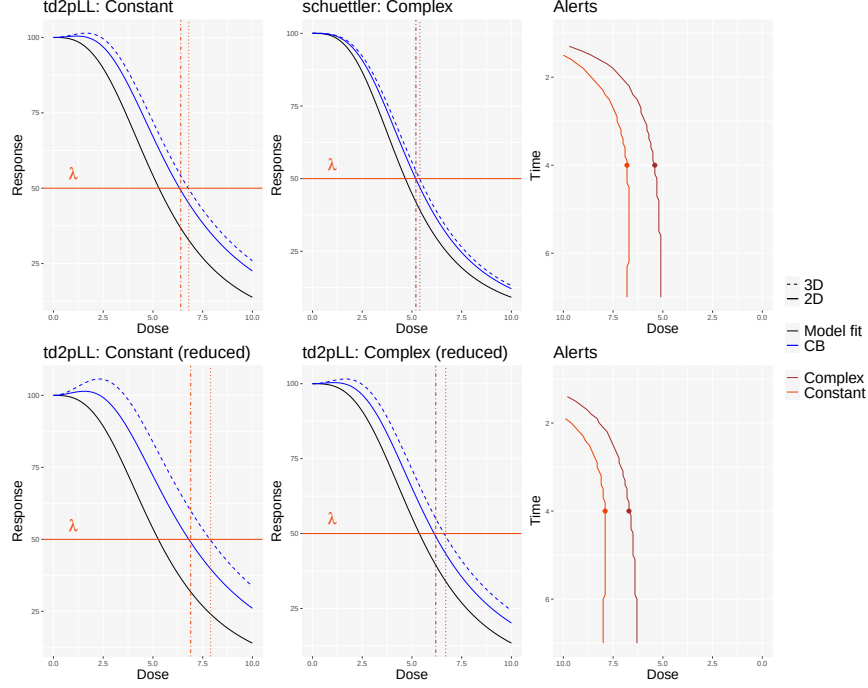


Figure 7: Visualization of the two-dimensional results at the fixed time point $\tilde{t} = 4$. The upper row shows the results for the full data set, the lower row for the artificially reduced data set. On the left, the estimated model fit and the confidence bands (CB) are displayed. These have been estimated from the two-dimensional (solid) and three-dimensional (dashed) approaches. The red horizontal line indicates the threshold $\lambda = 50\%$. The red and brown vertical lines highlight the respective identified alert dose. On the right, the estimated time-dose-alert relationship from the three-dimensional analysis is displayed. The dots indicate the alert for the fixed time point.

On the right, Figure 7 displays the two-dimensional time-dose-alert relationship, as identified by applying the three-dimensional approach. The difference between the constant model and the complex model is evident, with the doses identified for the complex case being smaller throughout the entire time span. The shape of the identified alert-relationship is consistent across all cases. For a brief period at the earliest time points, the maximum dose of 10 is too small to identify an alert. Subsequently, the identified alert decreases substantially up to approximately day 3. Thereafter, the decrease slows until the alert becomes almost constant around day 4. Therefore, the findings suggest that a shorter exposure period might have been sufficient for *ASP*.

6 Discussion

In this paper, a parametric methodology for the identification of alerts within a complex, three-dimensional data structure is proposed. This is achieved by focusing on a fixed time point and applying a two-dimensional hypothesis test, or by applying a three-dimensional hypothesis test to identify a time-dose-alert relationship. Thereby, we provide a tool that enables statistically meaningful interpretation of not only single alert doses, but

also of a two-dimensional relationship. As demonstrated by Scenario 1 in the simulation study, the integration of the GAMLSS framework in the bootstrap algorithm allows for a more precise analysis of complex standard deviation structures. The procedure can be applied to a broad range of statistical questions and a wide variety of models, as evidenced by the application of the *td2pLL*, *schuettler* and *Emax-Emax* models. As demonstrated by Scenario 2, in combination with an optimized design, the approach works stable even with a reduced number of observations. By introducing the fast approach, situations in which multiple runs are necessary and computation time is an issue is also considered.

Of course, this model has its limitations. As demonstrated by the simulation results in Section 4, there is a tendency to underestimate the true alert. The effect is relatively modest if the number of observations is not too small and the standard deviation structure is correctly specified. While the model appears relatively stable against misspecification of the standard deviation structure, the scope of the results presented in this manuscript may not be sufficient to provide a complete picture. Moreover, the approach is currently founded on model selection, particularly in regard to the parametric model selected for the mean. A more flexible approach to model design, such as model averaging (see, for example, Hagemann and Möllenhoff, 2025) or spline-based approaches (see Fahrmeir et al., 2022), might improve flexibility. Furthermore, the incorporation of random effects would be an interesting and important addition. Also, the hypothesis test was conducted using a simultaneous confidence band/plane, that was estimated based on a procedure that extended the approach of Möllenhoff et al. (2022). In principle, the hypothesis test may be conducted with any simultaneous confidence band/plane. A comparison of the various existing approaches may prove to be of interest.

In conclusion, we posit that our proposed method constitutes a valuable addition to the analysis of multidimensional dose-response data. It enables a comprehensive analysis of both two-dimensional and three-dimensional questions and is applicable to a wide range of complex data structures. The extension to more than three continuous variables is a straightforward process, thereby enhancing its applicability.

Supplementary material

The data presented in the case study are openly available in Archives of Toxicology at 10.1007/s00204-018-2302-0 or included in the R package *td2pLL*. Supplementary material providing additional tables and figures can be found online. Corresponding R code, which can be used to reproduce the analysis of the case study and the simulation results, is available at https://github.com/LuciaAmeis/Multivariate_alert_detection.

Funding

This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116).

References

- Ameis, L. and Möllenhoff, K. (2024). Identification of changes in gene expression. *arXiv preprint*.
- Arbo, M., Melega, S., Stöber, R., Schug, M., Rempel, E., Rahnenführer, J., Godoy, P., Reif, R., Cadenas, C., de Lourdes Bastos, M., et al. (2016). Hepatotoxicity of piperazine designer drugs: up-regulation of key enzymes of cholesterol and lipid biosynthesis. *Archives of Toxicology*, 90(12):3045–3060.
- Bornkamp, B., Pinheiro, J., and Bretz, F. (2009). Mcpmod: An r package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, 29:1–23.

- Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Bretz, F., Dette, H., and Pinheiro, J. (2010). Practical considerations for optimal designs in clinical dose finding studies. *Statistics in medicine*, 29(7-8):731–742.
- De Bastiani, F., Heller, G., Kneib, T., Mayr, A., Rigby, R., Stasinopoulos, M., et al. (2026). *GAMLSS*. Accessed February 05, 2026.
- Delignette-Muller, M.-L., Forfait, C., Billoir, E., and Charles, S. (2011). A new perspective on the Dunnett procedure: Filling the gap between NOEC/LOEC and ECx concepts. *Environmental Toxicology and Chemistry*, 30(12):2888–2891.
- Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.
- Duda, J., Drenda, C., Kästel, H., Rahnenführer, J., and Kappenberg, F. (2023). Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons. *Scientific Reports*, 13(1):20804.
- Duda, J., Hengstler, J., and Rahnenführer, J. (2022). td2ppl: An intuitive time-dose-response model for cytotoxicity data with varying exposure durations. *Computational Toxicology*, 23:100234.
- Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. CRC Press, 1st ed. edition.
- Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2022). *Regression: Models, Methods and Applications*. Springer Berlin, Heidelberg, 2nd edition.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.
- Galecki, A. and Burzykowski, T. (2012). Linear mixed-effects model. In *Linear mixed-effects models using R: a step-by-step approach*, pages 245–273. Springer.
- Ghallab, A., Cellière, G., Henkel, S., Driesch, D., Hoehme, S., Hofmann, U., Zellmer, S., Godoy, P., Sachinidis, A., Blaszkewicz, M., et al. (2016). Model-guided identification of a therapeutic strategy to reduce hyperammonemia in liver diseases. *Journal of hepatology*, 64(4):860–871.
- Gu, X., Albrecht, W., Edlund, K., Kappenberg, F., Rahnenführer, J., Leist, M., Moritz, W., Godoy, P., Cadenas, C., Marchan, R., et al. (2018). Relevance of the incubation period in cytotoxicity testing with primary human hepatocytes. *Archives of toxicology*, 92(12):3505–3515.
- Hagemann, N. and Möllenhoff, K. (2025). Overcoming model uncertainty—how equivalence tests can benefit from model averaging. *Statistics in Medicine*, 44(6):e10309.
- Hothorn, L. (2014). Statistical evaluation of toxicological bioassays - a review. *Toxicology Research*, 3(6):418–432.
- Huusari, R., Wang, T., Szedmak, S., Aittokallio, T., and Rousu, J. (2025). Predicting drug combination response surfaces. *npj Drug Discovery*, 2(1):2.
- ICCVAM (2001). *Guidance document on using in vitro data to estimate in vivo starting doses for acute toxicity*. Accessed March 18, 2026.
- Jensen, S., Kluxen, F., and Ritz, C. (2019). A review of recent advances in benchmark dose methodology. *Risk Analysis*, 39(10):2295–2315.

- Kappenberg, F., Duda, J., Schürmeyer, L., Gül, O., Brecklinghaus, T., Hengstler, J., Schorning, K., and Rahnenführer, J. (2023). Guidance for statistical design and analysis of toxicological dose–response experiments, based on a comprehensive literature review. *Archives of Toxicology*, 97(10):2741–2761.
- Kappenberg, F., Grinberg, M., Jiang, X., Kopp-Schneider, A., Hengstler, J., and Rahnenführer, J. (2021). Comparison of observation-based and model-based identification of alert concentrations from concentration-expression data. *Bioinformatics*, 37(14):1990 – 1996.
- Lange, M. and Schmidli, H. (2015). Analysis of clinical trials with biologics using dose–time-response models. *Statistics in Medicine*, 34(22):3017–3028.
- Lee, S. (2010). Drug interaction: focusing on response surface models. *Korean journal of anesthesiology*, 58(5):421–434.
- Loewe, S. (1953). The problem of synergism and antagonism of combined drugs. *Arzneimittel-forschung*, 3(6):285–290.
- Manning, C. (2008). *Introduction to information retrieval*. Syngress Publishing,.
- Martin-Betancor, K., Ritz, C., Fernández-Piñas, F., Leganés, F., and Rodea-Palomares, I. (2015). Defining an additivity framework for mixture research in inducible whole-cell biosensors. *Scientific Reports*, 5(1):17200.
- Mokhtari, R., Homayouni, T., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017). Combination therapy in combating cancer. *Oncotarget*, 8(23):38022.
- Möllenhoff, K., Schorning, K., and Kappenberg, F. (2022). Identifying alert concentrations using a model-based bootstrap approach. *Biometrics*, 79(3):2076–2088.
- Riddell, R., Panacer, D., Wilde, S., Clothier, R., and Balls, M. (1986). The importance of exposure period and cell type in in vitro cytotoxicity tests. *Alternatives to Laboratory Animals*, 14(2):86–92.
- Rigby, R. and Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3):507–554.
- Schürmeyer, L., Sandig, L., Hezler, L., Igl, B.-W., and Schorning, K. (2025). Optimal designs for identifying effective doses in drug combination studies. *arXiv preprint*.
- Schüttler, A., Altenburger, R., Ammar, M., Bader-Blukott, M., Jakobs, G., Knapp, J., Krüger, J., Reiche, K., Wu, G., and Busch, W. (2019). Map and model—moving from observation to prediction in toxicogenomics. *GigaScience*, 8(6):giz057.
- Sebaugh, J. L. (2011). Guidelines for accurate ec50/ic50 estimation. *Pharmaceutical Statistics*, 10(2):128–134.
- Stasinopoulos, D. and Rigby, R. (2008). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46.
- Stasinopoulos, M., Rigby, R., Heller, G., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman and Hall/CRC.
- Zhou, Y., Sloan, A., Menon, S., and Wang, L. (2025). Combination mcp-mod for two-drug combination dose-ranging studies. *Journal of Biopharmaceutical Statistics*, 35(2):257–270.

Supplement: Estimating effect thresholds and beyond: A flexible framework for multivariate alert detection

L. Ameis^{1,2}, N. Hagemann^{1,2,3}, and K. Möllenhoff^{1,2*}

¹Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine,
University of Cologne

²Division of Mathematics, Department of Mathematics and Computer Science, University of
Cologne

³Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University

2 Parametric models used in the paper

In the following section, we present the three parametric models used in this paper.

td2pLL

Here, the model introduced in Duda et al. (2022) for analyzing cytotoxicity data is presented. It is based on the *4pLL/sigEmax* model (Bornkamp et al., 2009; Bornkamp, B. and Pinheiro, J. and Bretz, F., 2017) with fixed lower and upper asymptotes. The EC_{50} parameter is modeled as time-dependent by applying a modified version of Haber's law (Haber, 1924; Miller et al., 2000).

- C_0 The limit EC_{50} for a hypothetical infinite t
- $\gamma > 0$ Indicates when the time dependent $EC_{50}(t) := C_0 + \Delta t^{-\gamma}$ changes most between different time points
 - Larger γ indicates large changes between early time points
 - Smaller γ indicates large changes between larger time points
 - Vice versa for a small γ
- $\gamma < 0$ A beneficial effect is observed (usually irrelevant for cytotoxicity data)
- Δ Difference between $EC_{50}(1)$ and C_0
 - Indicates the range of $EC_{50}(t)$ values for $t \geq 1$
- h Slope parameter, similar to *4pLL* (Bornkamp et al., 2009)

$$f(t, d, (C_0, \gamma, \Delta, h)) = 100 - 100 \frac{d^h}{d^h + (C_0 + \Delta t^{-\gamma})^h}$$

schuettler

Here, the model introduced in Schüttler et al. (2019) for toxicogenomics is presented. It is based on the *4pLL/sigEmax* model, similar to the *td2pLL* model presented above. However, the EC_{50} parameter is modeled by a logarithmic Gaussian function.

- $\log FC_{\max}$ Maximum change (of $\log FC$ in original paper, here stand-in for maximum change in general)
- slope* slope parameter
- $S_{\max}$ Maximum of $\frac{1}{EC_{50}(t)} = S_{\max} \exp\left(-0.5 \left(\frac{\log(t) - \log(t_{\max})}{S_{\text{dur}}}\right)^2\right)$
- $t_{\max}$ Point in time with $S_{\max}$
- S_{dur} Measure of duration of the $EC_{50}(t)$ interval

$$f(t, d, (\log FC_{\max}, \text{slope}, S_{\max}, t_{\max}, S_{\text{dur}})) = \frac{\log FC_{\max}}{1 + \exp\left(-\text{slope} \left(\log(d) - \log\left(\frac{1}{S_{\max} \exp\left(-0.5 \left(\frac{\log(t) - \log(t_{\max})}{S_{\text{dur}}}\right)^2}\right)\right)\right)\right)}$$

E_{max}-E_{max}

Here, a drug combination model based on the *E_{max}* model is presented. In general, let two substances, denoted by D_1 and D_2 , be modeled by the individual regression functions $\eta_{D_1}(d_1, \boldsymbol{\theta}_{D_1})$ and $\eta_{D_2}(d_2, \boldsymbol{\theta}_{D_2})$, respectively. The combination of these two drugs can then be modeled by

$$f((d_1, d_2), (\theta_0, \boldsymbol{\theta}_{D_1}, \boldsymbol{\theta}_{D_2}, \gamma)) = \theta_0 + \eta_{D_1}(d_1, \boldsymbol{\theta}_{D_1}) + \eta_{D_2}(d_2, \boldsymbol{\theta}_{D_2}) + \gamma \eta_{D_1}(d_1, \boldsymbol{\theta}_{D_1}) \eta_{D_2}(d_2, \boldsymbol{\theta}_{D_2}).$$

Note that the individual placebo parameters of η_{D_1} and η_{D_2} are removed and combined in the joined parameter θ_0 . The parameter γ indicates the interaction effect of the two drugs. Precisely, the interaction effect is positive if $\gamma > 0$ and negative if $\gamma < 0$. If $\gamma = 0$, no interaction between the two substances is observed.

The drug combination model presented in this manuscript is based on two *E_{max}* models. The *E_{max}* model as presented by Bornkamp et al. (2009) is parameterized as follows:

$E_{\max}$ Maximum effect
 ED_{50} The dose/time point, which produces half of $E_{\max}$

$$\eta_{Emax}(d, (E_{\max}, ED_{50})) = E_{\max} \frac{d}{ED_{50} + d}$$

Note that the placebo parameter was removed from the parametrization.

References

- Bornkamp, B., Pinheiro, J., and Bretz, F. (2009). Mcpmod: An r package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, 29:1–23.
- Bornkamp, B. and Pinheiro, J. and Bretz, F. (2017). *MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies*. R package version 1.0-10.1.
- Duda, J., Hengstler, J., and Rahnenführer, J. (2022). td2pll: An intuitive time-dose-response model for cytotoxicity data with varying exposure durations. *Computational Toxicology*, 23:100234.
- Haber, F. (1924). *Zur Geschichte des Gaskrieges*, pages 76–92. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Miller, F., Schlosser, P., and Janszen, D. (2000). Haber’s rule: a special case in a family of curves relating concentration and duration of exposure to a fixed level of response for a given endpoint. *Toxicology*, 149(1):21–34.
- Schüttler, A., Altenburger, R., Ammar, M., Bader-Blukott, M., Jakobs, G., Knapp, J., Krüger, J., Reiche, K., Wu, G., and Busch, W. (2019). Map and model—moving from observation to prediction in toxicogenomics. *GigaScience*, 8(6):giz057.

4 Simulation study

4.1 Setting and data generation

Table A: Design details of the simulated scenarios.

Scenario	Design	Measurements $((t_p, d_q), n_{pq}) / ((d_{1,p}, d_{2,q}), n_{pq})$
1	Full	$((1, 0), 24); ((1, 0.1), 12); ((1, 0.316), 12); ((1, 1), 12); ((1, 3.16), 12);$ $((1, 10), 12); ((2, 0), 24); ((2, 0.1), 12); ((2, 0.316), 12); ((2, 1), 11);$ $((2, 3.16), 11); ((2, 10), 12); ((7, 0), 24); ((7, 0.1), 12); ((7, 0.316), 12);$ $((7, 1), 12); ((7, 3.16), 12); ((7, 10), 12)$
1	Reduced	$((1, 0), 3); ((1, 0.1), 3); ((1, 0.316), 3); ((1, 1), 3); ((1, 3.16), 3);$ $((1, 10), 3); ((2, 0), 3); ((2, 0.1), 3); ((2, 0.316), 3); ((2, 1), 3);$ $((2, 3.16), 3); ((2, 10), 3); ((7, 0), 3); ((7, 0.1), 3); ((7, 0.316), 3);$ $((7, 1), 3); ((7, 3.16), 3); ((7, 10), 3)$
2	3x3 Factorial	$((0, 0), 17); ((0, 6), 17); ((0, 12), 17); ((5, 0), 17); ((5, 6), 17);$ $((5, 12), 17); ((10, 0), 17); ((10, 6), 17); ((10, 12), 16)$
2	3x3 Factorial	$((0, 0), 10); ((0, 6), 10); ((0, 12), 10); ((5, 0), 10); ((5, 6), 10);$ $((5, 12), 10); ((10, 0), 10); ((10, 6), 10); ((10, 12), 10)$
2	3x3 Factorial	$((0, 0), 5); ((0, 6), 5); ((0, 12), 5); ((5, 0), 5); ((5, 6), 5);$ $((5, 12), 5); ((10, 0), 5); ((10, 6), 5); ((10, 12), 5)$
2	D-optimal	$((0, 0), 25); ((0, 12), 25); ((1.936057, 12), 22); ((2.147924, 4.218333), 10);$ $((10, 0), 25); ((10, 3.852853), 21); ((10, 12), 24)$
2	D-optimal	$((0, 0), 14); ((0, 12), 15); ((1.936057, 12), 13); ((2.147924, 4.218333), 6); ((10, 0), 15);$ $((10, 3.852853), 13); ((10, 12), 14)$
2	D-optimal	$((0, 0), 14); ((0, 12), 15); ((1.936057, 12), 13); ((2.147924, 4.218333), 6); ((10, 0), 15);$ $((10, 3.852853), 13); ((10, 12), 14)$

4.2 Results for two-dimensional hypothesis

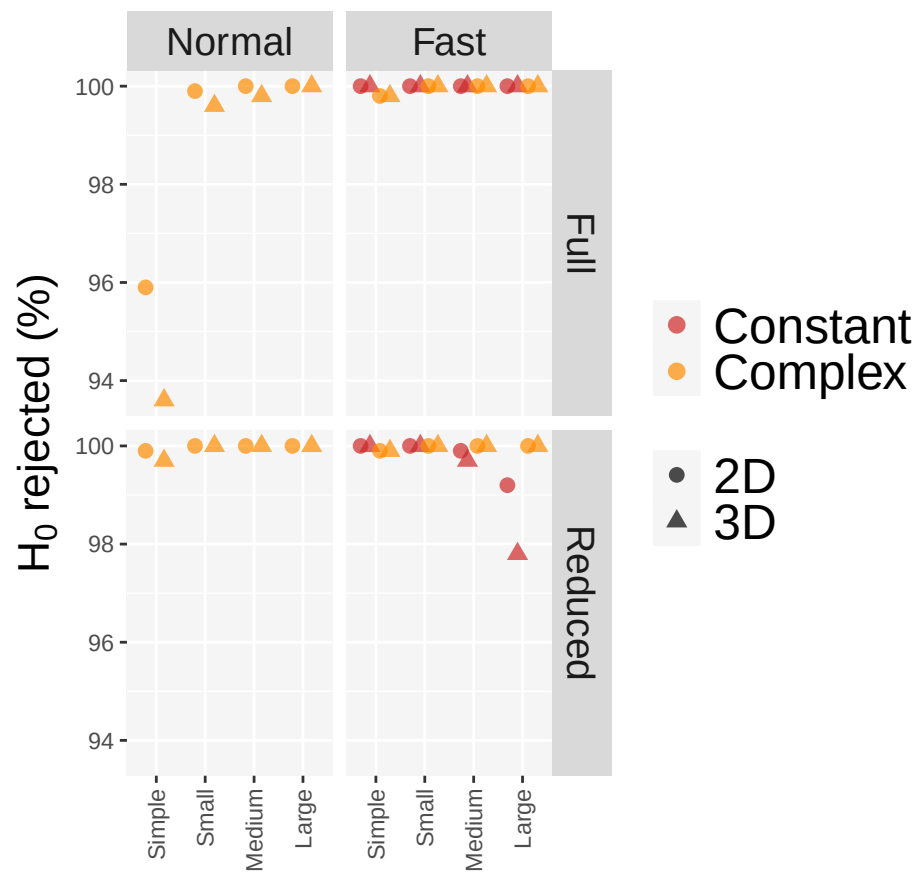


Figure A: Number of rejected H_0 in percent for Scenario 1 including the normal and fast algorithms.

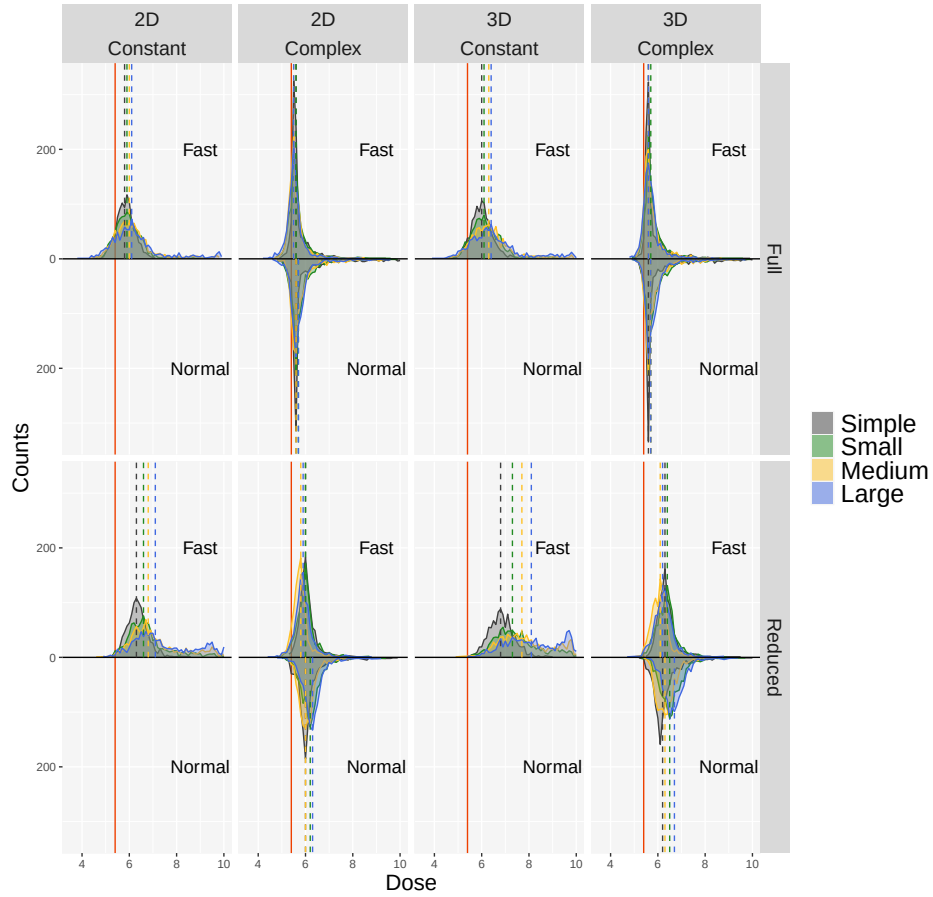


Figure B: Counts of identified alerts at different dose levels for the two-dimensional approach. Each block contains the results of a model assumption regarding σ (constant or complex) for the full or reduced design, respectively. The colors differentiate between the various simulation scenarios with simple, small, medium and large standard deviation. The red vertical line indicates the true alert. The dashed vertical lines show the respective medians. The results for the fast algorithm are depicted above the 0 line, the results for the normal algorithm for comparison below the 0 line.

Table B: Simulation results for Scenario 1 with the full design for the two-dimensional hypothesis.

	Approach	Algorithm	Assumption	Simple	Small	Medium	Large
H_0 rejected (%)	2D	Fast	Constant	100	100	100	100
	3D	Fast	Constant	100	100	100	100
Median	2D	Fast	Constant	5.8	5.9	6.0	6.1
	3D	Fast	Constant	6.0	6.1	6.3	6.4
Standard deviation	2D	Fast	Constant	0.387	0.597	0.851	1.109
	3D	Fast	Constant	0.424	0.651	0.913	1.152
H_0 rejected (%)	2D	Normal	Complex	95.9	99.9	100	100
	3D	Normal	Complex	93.6	99.6	998	1000
	2D	Fast	Complex	99.8	100	100	100
	3D	Fast	Complex	99.8	100	100	100
Median	2D	Normal	Complex	5.6	5.7	5.6	5.7
	3D	Normal	Complex	5.6	5.7	5.7	5.7
	2D	Fast	Complex	5.6	5.6	5.5	5.5
	3D	Fast	Complex	5.6	5.7	5.6	5.6
Standard deviation	2D	Normal	Complex	0.882	0.645	0.669	0.600
	3D	Normal	Complex	0.892	0.654	0.685	0.598
	2D	Fast	Complex	0.486	0.494	0.440	0.446
	3D	Fast	Complex	0.547	0.513	0.473	0.448

Table C: Simulation results for Scenario 1 with reduced design for the two-dimensional hypothesis.

	Approach	Algorithm	Assumption	Simple	Small	Medium	Large
H_0 rejected (%)	2D	Fast	Constant	100.0	100.0	99.9	99.2
	3D	Fast	Constant	100.0	100.0	99.7	97.8
Median	2D	Fast	Constant	6.3	6.6	6.8	7.1
	3D	Fast	Constant	6.8	7.3	7.7	8.1
Standard deviation	2D	Fast	Constant	0.443	0.914	1.099	1.248
	3D	Fast	Constant	0.542	0.957	1.057	1.121
H_0 rejected (%)	2D	Normal	Complex	99.9	100.0	100.0	100.0
	2D	Normal	Complex	99.7	100.0	100.0	100.0
	2D	Fast	Complex	99.9	100.0	100.0	100.0
	3D	Fast	Complex	99.9	100.0	100.0	100.0
Median	2D	Normal	Complex	6.0	6.2	6.0	6.3
	3D	Normal	Complex	6.2	6.5	6.3	6.7
	2D	Fast	Complex	6.0	6.0	5.8	5.9
	3D	Fast	Complex	6.3	6.4	6.1	6.2
Standard deviation	2D	Normal	Complex	0.557	0.494	0.398	0.498
	3D	Normal	Complex	0.606	0.501	0.419	0.488
	2D	Fast	Complex	0.372	0.434	0.325	0.438
	3D	Fast	Complex	0.409	0.453	0.36	0.448

Table D: Simulation results for Scenario 2 with the Factorial 3×3 design for the two-dimensional hypothesis

	Approach	N152	N90	N45
H_0 rejected (%)	2D	99.6	93.1	63.9
	3D	98.0	79.6	38.6
Median	2D	9	9.4	9.9
	3D	9.4	9.9	10.4
Standard deviation	2D	0.597	0.869	0.971
	3D	0.644	0.851	0.874

Table E: Simulation results for Scenario 2 with the D-optimal design for the two-dimensional hypothesis.

	Approach	N152	N90	N45
H_0 rejected (%)	2D	99.8	93.6	94.7
	3D	98.9	82.7	81.9
Median	2D	9.1	9.6	9.6
	3D	9.5	10.1	10.1
Standard deviation	2D	0.654	0.881	0.898
	3D	0.703	0.866	0.892

4.3 Results for three-dimensional hypothesis

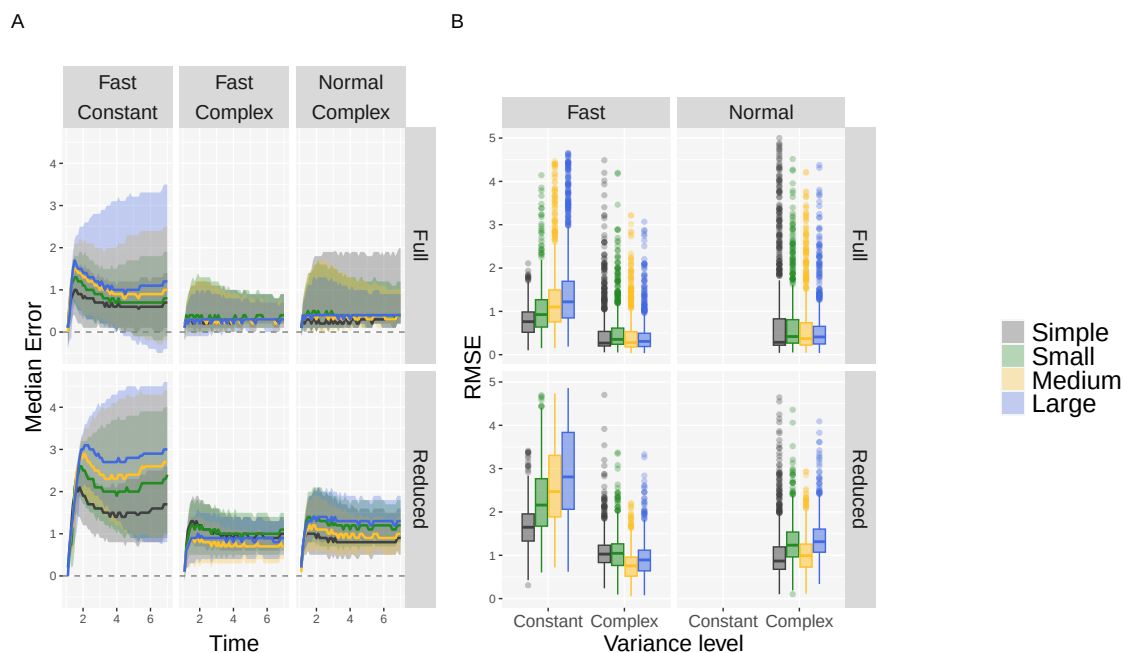


Figure C: (A) Median error of the identified alert doses along the time-axis for the three-dimensional approach. Each block contains the results of a model assumption regarding σ (constant or complex) for the full or reduced design, respectively. The solid line indicates the median. The areas indicate the 10% to 90% quantiles. (B) Boxplots of the RMSEs for the three-dimensional approach.

Table F: Simulation results for Scenario 1 with full design for the three-dimensional hypothesis.

	Algorithm	Assumption	Simple	Small	Medium	Large
H_0 rejected (%)	Fast	Constant	100	100	100	100
	Normal	Complex	95.3	99.7	99.9	100
	Fast	Complex	99.9	100	100	100
Mean Recall (%)	Fast	Constant	97.4	96.5	95.8	94.9
	Normal	Complex	95.2	97.6	97.3	98.1
	Fast	Complex	97.6	98.3	98.5	98.8
Standard deviation Recall (%)	Fast	Constant	1.7	1.5	1.8	2.6
	Normal	Complex	10.5	4.1	4.9	2.9
	Fast	Complex	3.6	1.5	.16	1.0
Mean Precision (%)	Fast	Constant	100	100	100	100
	Normal	Complex	100	100	100	100
	Fast	Complex	100	100	100	100
Standard deviation Precision (%)	Fast	Constant	0.1	0	0	0
	Normal	Complex	0	0.1	0	0
	Fast	Complex	0	0.1	0	0
Mean Onset Error	Fast	Constant	0.156	0.212	0.253	0.307
	Normal	Complex	0.285	0.140	0.154	0.113
	Fast	Complex	0.142	0.099	0.091	0.073
Standard deviation Onset Error	Fast	Constant	0.101	0.093	0.110	0.159
	Normal	Complex	0.615	0.209	0.244	0.175
	Fast	Complex	0.215	0.093	0.093	0.061
Mean Offset Error	Fast	Constant	0	0	0	0
	Normal	Complex	-0.003	-0.003	-0.005	0
	Fast	Complex	0	0	0	0
Standard deviation Offset Error	Fast	Constant	0	0	0	0
	Normal	Constant	0.1	0.108	0.127	0
	Fast	Constant	0	0	0	0

Table G: Simulation results for Scenario 1 with reduced design for the three-dimensional hypothesis.

	Algorithm	Assumption	Simple	Small	Medium	Large
H_0 rejected (%)	Fast	Constant	100	100	99.7	98.0
	Normal	Complex	99.9	100	100	100
	Fast	Complex	100	100	100	100
Mean Recall (%)	Fast	Constant	93.0	90.1	88.5	0.863
	Normal	Complex	95.8	96.0	97.0	0.968
	Fast	Complex	96.0	97.0	97.9	98.0
Standard deviation Recall (%)	Fast	Constant	2.3	3.5	3.6	5.7
	Normal	Complex	4.8	2.3	1.6	1.8
	Fast	Complex	3.2	1.6	1.1	1.1
Mean Precision (%)	Fast	Constant	100	100	100	100
	Normal	Complex	100	100	100	100
	Fast	Complex	100	100	100	100
Standard deviation Precision (%)	Fast	Constant	0	0	0	0
	Normal	Complex	0	0	0	0
	Fast	Complex	0	0	0	0
Mean Onset Error	Fast	Constant	0.422	0.592	0.689	0.814
	Normal	Complex	0.253	0.238	0.181	0.194
	Fast	Complex	0.239	0.180	0.126	0.118
Standard deviation Onset Error	Fast	Constant	0.137	0.171	0.198	0.291
	Normal	Complex	0.287	0.136	0.095	0.106
	Fast	Complex	0.191	0.095	0.068	0.068
Mean Offset Error	Fast	Constant	0	-0.002	-0.002	-0.006
	Normal	Complex	0	0	0	0
	Fast	Complex	0	0	0	0
Standard deviation Offset Error	Fast	Constant	0	0.076	0.051	0.11
	Normal	Complex	0	0	0	0
	Fast	Complex	0	0	0	0

Table H: Simulation results for Scenario 2 with factorial 3×3 design for the three-dimensional hypothesis.

	N152	N90	N45
H_0 rejected (%)	99.6	94.1	59.7
Mean Recall (%)	69.6	60.7	52.8
Standard deviation Recall (%)	9.3	14.7	17.8
Mean Precision (%)	100	100	100
Standard deviation Precision (%)	0.1	0.1	0.5
Mean Onset Error	1.979	2.553	3.019
Standard deviation Onset Error	0.605	0.955	1.137
Mean Offset Error	0.000	-0.004	-0.049
Standard deviation Offset Error	0.000	0.089	0.270

Table I: Simulation results for Scenario 2 with D-otptimal design for the three-dimensional hypothesis.

	N152	N90	N45
H_0 rejected (%)	100	96.4	98.0
Mean Recall (%)	74.1	62.5	62.1
Standard deviation Recall (%)	10.0	16.9	16.5
Mean Precision (%)	100	100	100
Standard deviation Precision (%)	0.1	0	0.1
Mean Onset Error	1.683	2.436	2.465
Standard deviation Onset Error	0.651	1.099	1.070
Mean Offset Error	0.000	0.000	0.000
Standard deviation Offset Error	0.000	0.000	0.000

5 Case Study

5.1 Comparison of the GAMLSS fits to the standard deviation

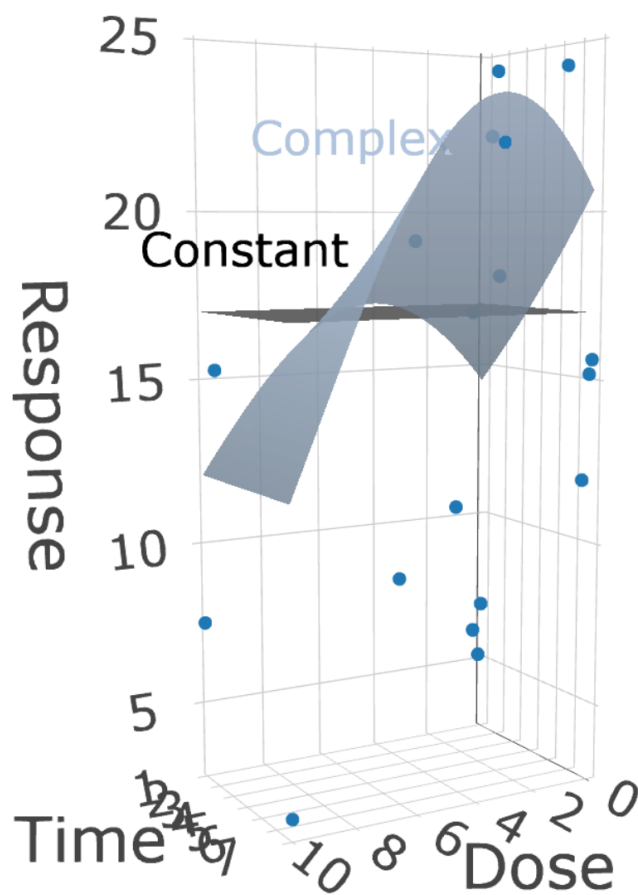


Figure D: Comparison of the GAMLSS fits to the standard deviation for the three different cases when the *td2pLL* model is assumed for the model fit. The blue dots indicate the standard deviation for each combination of measured dose and time.

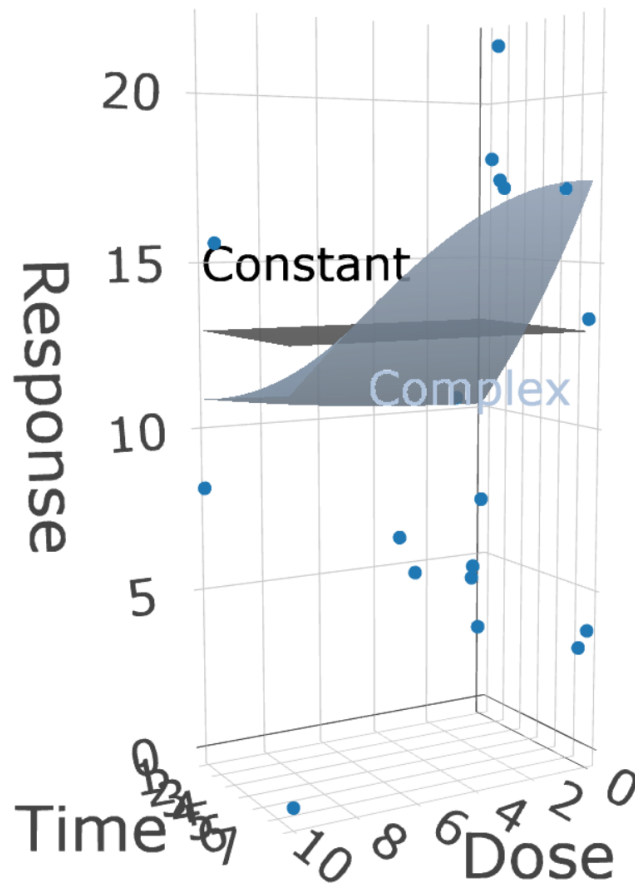


Figure E: Comparison of the GAMLSS fits to the standard deviation for the two different cases when the *td2pLL* model is assumed for the model fit and the data is reduced. The blue dots indicate the standard deviation for each combination of measured dose and time.

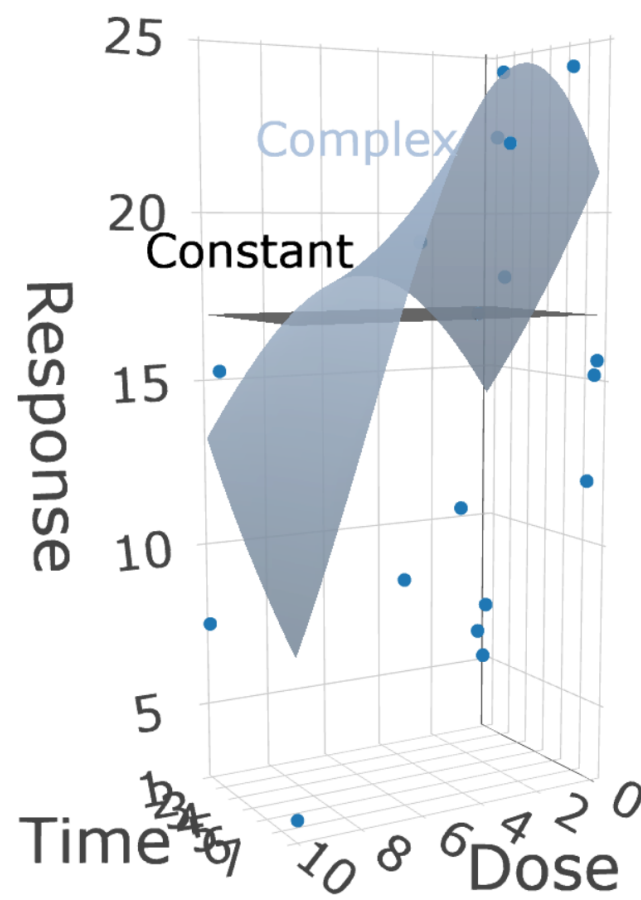


Figure F: Comparison of the GAMLSS fits to the standard deviation for the two different cases when the *schuettler* model is assumed for the model fit. The blue dots indicate the standard deviation for each combination of measured dose and time.

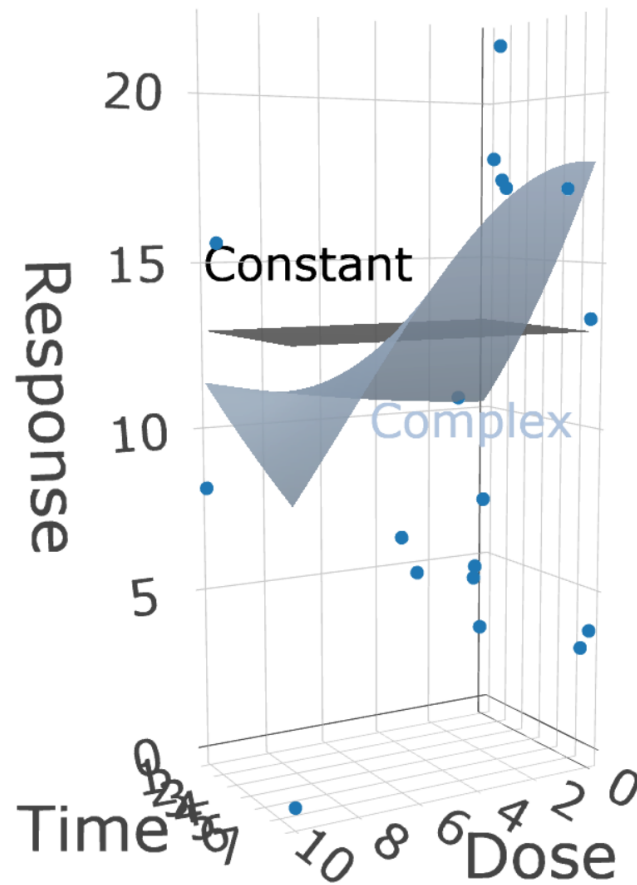


Figure G: Comparison of the GAMLSS fits to the standard deviation for the two different cases when the *schuettler* model is assumed for the model fit and the data is reduced. The blue dots indicate the standard deviation for each combination of measured dose and time.

5.2 Three dimensional results (Larger AIC)

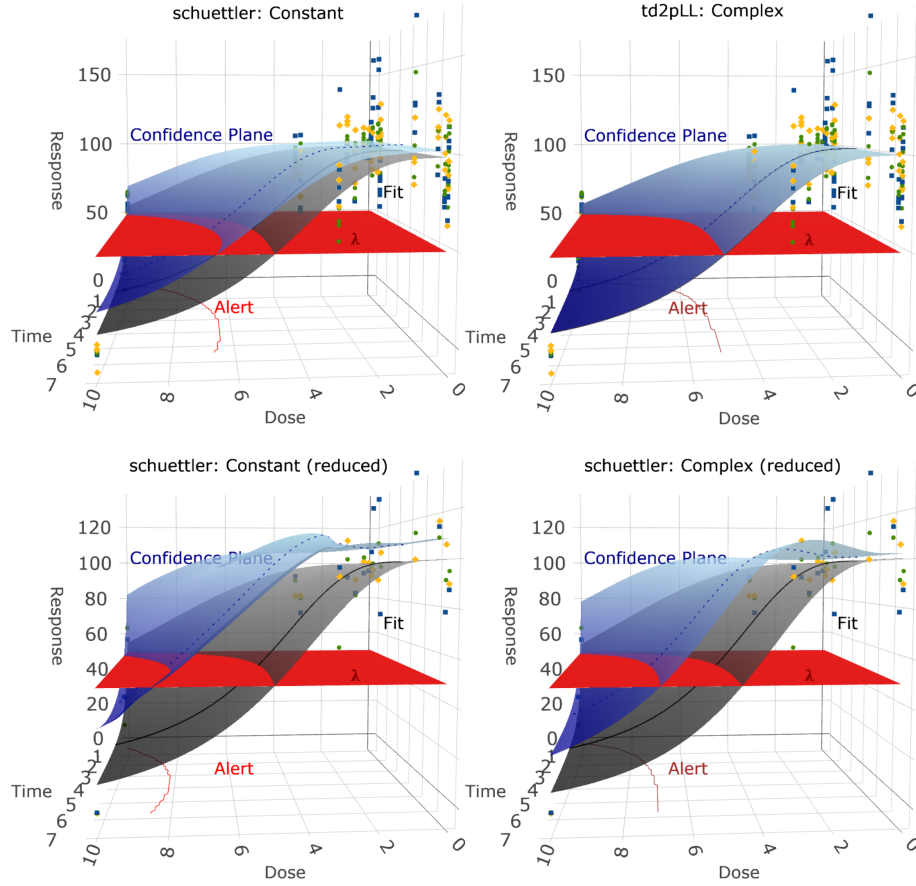


Figure H: Visualization of the results obtained from the three-dimensional approach when fitting the alternative models. The upper row displays the results for the full data set, the lower row the results for the artificially reduced data set. The dots represent the measured data points, which have been colored according to the donor. The black plane indicates the model fit. The blue plane represents the estimated confidence plane. The fixed time point $t_0 = 4$ is highlighted in both planes. The alert threshold λ is displayed as a red hyperplane that is perpendicular to the *Response*-axis. The estimated alert relationship is projected onto the plane spanned by the *Dose*-axis and the *Time*-axis intersecting the *Response*-axis at the origin.

5.3 Two dimensional results

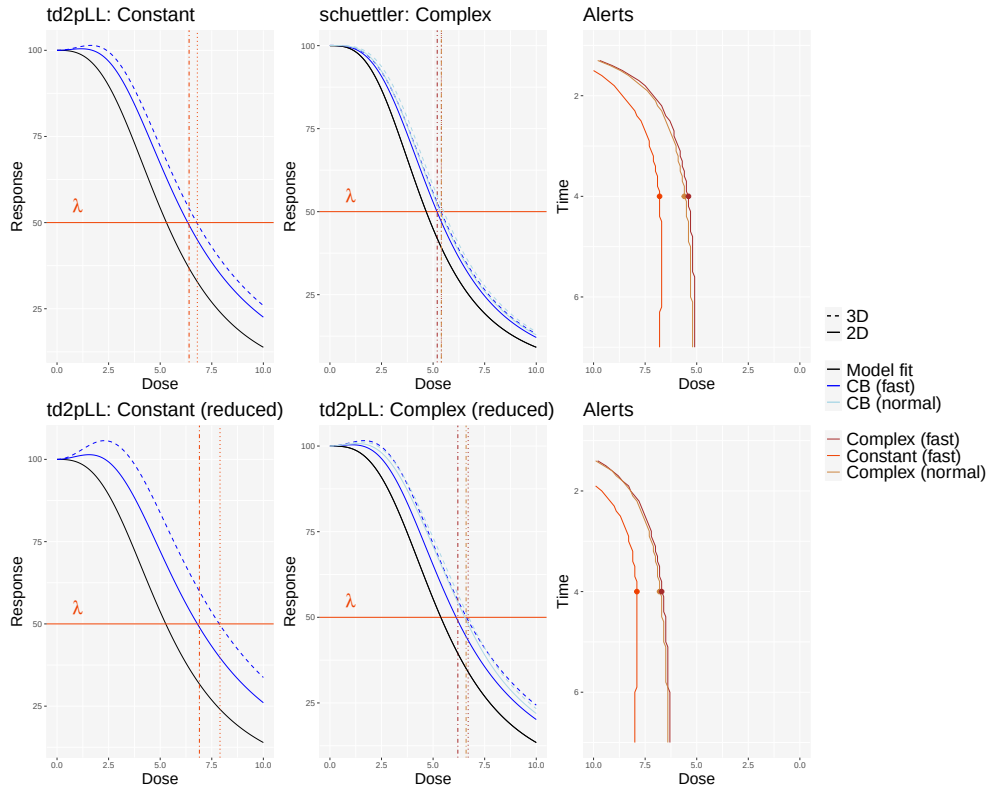


Figure I: Visualization of the two-dimensional results at the fixed time point $\tilde{t} = 4$ including results from the fast and the normal algorithm. The upper row shows the results for the full data set, the lower row for the artificially reduced data set. On the left, the estimated model fit and the confidence bands (CB) are displayed. These have been estimated from the two-dimensional (solid) and three-dimensional (dashed) approaches by applying the normal (light-blue) or the fast (blue) algorithm. The red horizontal line indicates the threshold $\lambda = 50\%$. The red and orange/light-brown vertical lines highlight the respective identified alert dose. On the right, the alert relationship between *Dose* and *Time* estimated from the three dimensional analysis is displayed. The dots indicate the alert for the fixed time point.

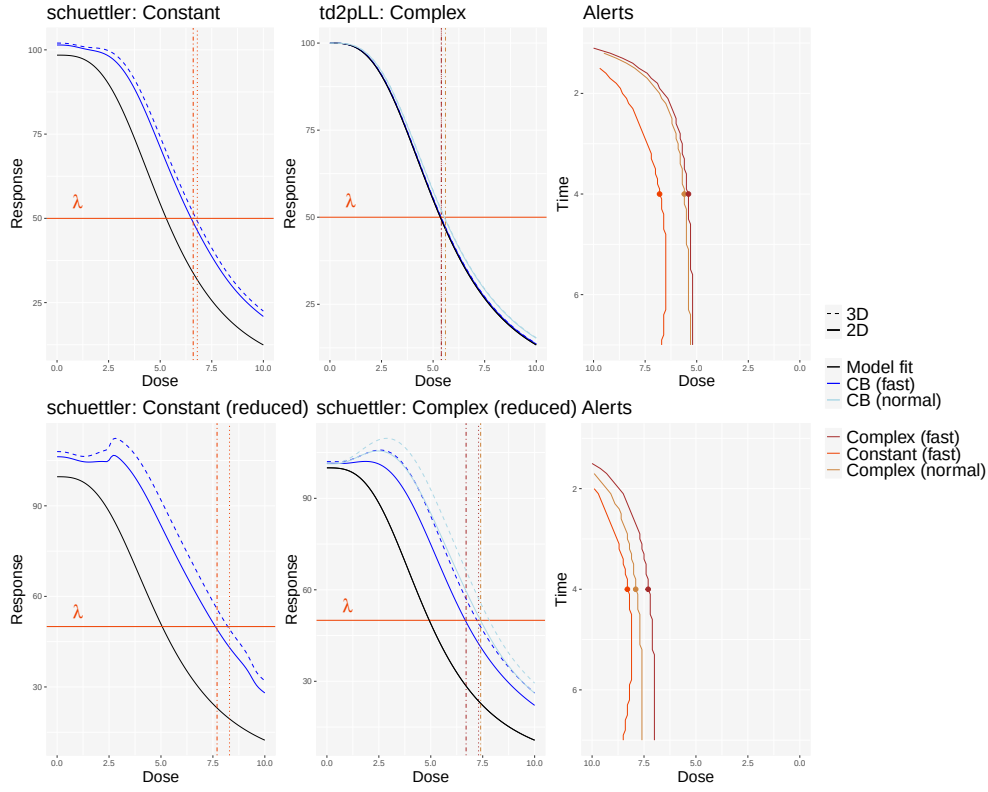


Figure J: Visualization of the two-dimensional results at the fixed time point $\tilde{t} = 4$ for the alternative model fits including results from the fast and the normal algorithm. The upper row shows the results for the full data set, the lower row for the artificially reduced data set. On the left, the estimated model fit and the confidence bands (CB) are displayed. These have been estimated from the two-dimensional (solid) and three-dimensional (dashed) approaches by applying the normal (light-blue) or the fast (blue) algorithm. The red horizontal line indicates the threshold $\lambda = 50\%$. The red and orange/light-brown vertical lines highlight the respective identified alert dose. On the right, the alert relationship between *Dose* and *Time* estimated from the three-dimensional analysis is displayed. The dots indicate the alert for the fixed time point.

5.4 Results in Tables

Table J: All results for the full data set.

Model	Case	Algorithm	2D alert	3D alert
td2pLL	Case 1	fast	6.4	6.8
td2pLL	Case 2	normal	5.6	5.6
td2pLL	Case 2	fast	5.4	5.4
schuettler	Case 1	fast	6.6	6.8
schuettler	Case 2	normal	5.4	5.6
schuettler	Case 2	fast	5.2	5.4

Table K: All results for the reduced data set

Model	Case	Algorithm	2D alert	3D alert
td2pLL	Case 1	Fast	6.9	7.9
td2pLL	Case 2	Normal	6.6	6.8
td2pLL	Case 2	Fast	6.2	6.7
schuettler	Case 1	Fast	7.7	8.3
schuettler	Case 2	Normal	7.4	7.9
schuettler	Case 2	Fast	6.7	7.3

4 Discussion

The three articles presented in this thesis discuss the interplay between research question, data structure, and methodology. Two thematic areas are investigated: risk estimation in time-to-event analysis and model-based alert identification. The first article discusses the former, while the second and third article examine the latter.

In the first article, a nonparametric PR estimator for the RR of two groups is proposed. This aligns the research question of by what factor the risk for one group is higher than for the other, hence circumventing conceptional problems of the HR and OR, as well as the potential for misinterpretation. As the method is based on the Kaplan-Meier estimator, it avoids restrictive parametric assumptions and is therefore broadly applicable. An extensive simulation study reveals that the proposed estimator mostly outperforms a parametric alternative in terms of precision and numerical robustness. Additionally, the simulation study compares the results of the proposed approach to those of the Cox's PH model and the log-logistic PO model when their respective assumptions are fulfilled and the HR, OR and RR are used interchangeably. However, since the proposed estimator calculates the mean RR over time, it can be meaningfully interpreted in most cases, even if the assumption of PR is violated. The results clearly emphasize the importance of differentiating the three values. A bootstrap approach can be utilized to construct confidence intervals for the estimator and test for significant group differences.

Naturally, this approach come with some limitations and avenues for further research both of which are mentioned in the article. For proposed approach, two main topics warrant further exploration: the inclusion of more than one covariate and an investigation of the PR assumption. A procedure that is applicable to higher-dimensional settings can include more information, thereby enhancing the results. A potential extension to more categorical covariates could be based on further stratification, but might result in subgroups with a small number of events. Moreover, the incorporation of continuous covariates constitutes an important extension. With regard to the assumption of PR, this is not always fulfilled. However, in instances where the CDFs cross, the mean RR might be a misleading indicator. Inspired by recommendations for the mean HR (Hernán 2010), this could be accommodated for by applying the proposed procedure to individual time intervals. Nonetheless, the procedure would benefit from a goodness-of-fit test for the PR assumption.

Turning to the second thematic area of this thesis, the second article proposes a model-based approach to estimate the full covariate region, e.g., time frame, during which the response changes. This provides a tool that extends the interpretation beyond the mere onset of an effect. The procedure is based on a hypothesis test and produces statistically significant results at the α -level. An extensive case study demonstrates a plethora of possible applications to gene expression data. Examples of applicable models include the 4pLL model, the beta model, the exponential model, the quadratic model and B-

splines, which illustrate the broad range of response shapes that can be accommodated. In addition to the most basic application of the proposed approach, an application is demonstrated that compares the time frames for the same gene in two groups. A third application combines the proposed approach with a gene ontology enrichment analysis (Alexa et al. 2006; Alexa et al. 2024), which allows inference to be made about the underlying biological processes. This clearly illustrates the benefits of shifting the research question.

In the third article, the concept of model-based alert identification is extended to three-dimensional data structures in a manner that naturally generalizes to higher-dimensional settings. Similar to the approach introduced in the second article, the discussed method is based on a hypothesis test and thus ensures inference based on statistical significance. The formulation of the hypothesis can be adapted to the underlying research question: to identify an alert in one covariate while the other is fixed or to identify an alert relationship between covariates. The integration of GAMLSS into the initial model fit and the bootstrap sample generation ensures that the procedure can accommodate the complex nature of multidimensional data. The clear benefit of this procedure is investigated in the case study and in Scenario 1 of the simulation study.

Both methods come with limitations and possibilities for further research, which are discussed in the articles. One such limitation is a clear tendency for the procedures to underestimate the alert or covariate region. This effect is modest in case of lower data variability but becomes more pronounced the higher the variability is. This outcome was anticipated for two primary reasons: The approaches are significance-based, which has been demonstrated to yield results that are often conservative (Bretz et al. 2005; Kappenberg et al. 2021). Additionally, the approaches are based on simultaneous inference, which was similarly linked to conservative results (see, e.g., Wu et al. (2006), Deutsch et al. (2013), and Deutsch et al. (2012)). As discussed above, both properties are important, depending on the research question, and should be included in the approach if necessary. However, the methods would benefit from an extension that combats this effect. One potential solution to this issue would be to relax the significance condition in a manner similar to Bretz et al. (2005). Additionally, the hypothesis tests in both articles can in theory be conducted utilizing any simultaneous confidence band or plane for the model or the first derivative of the model, respectively. Adaptions of alternative construction methods might show more stable results. Also, it is important to note that both procedures rely on model fitting, decided on by an initial model selection step. This could lead to model-misspecification. A more flexible choice of model design, such as model-averaging (see, e.g., Lin et al. (2012), Kang et al. (2000), Jensen et al. (2015), Jensen et al. (2019), and Hagemann et al. (2025)) could lead to improvements. With regard to the second article, an extension of the method to a higher-dimensional setting similar to the approach presented in the third article would be of interest. This extension would serve to further expand the applicability of the method. With regard to the third article, the results obtained by incorporating spline-based models analogous to the second article

would be of particular interest. As with potential model-averaging extensions, this has the capacity to enhance adaptability to diverse data structures.

In summary, three methods are developed that provide advances and improve the alignment between current research questions and statistical inference. They also account for, or extend to, different underlying data structures. Performance is evaluated by simulation studies presented in each of the articles. In addition, case studies from diabetes research or toxicology demonstrate the practical relevance of each of the methods.

References

- Akkaya Hocagil, T., Ryan, L. M., Cook, R. J., Dang, K.-D., Carter, R. C., Richardson, G. A., Day, N. L., Coles, C. D., Carmichael Olson, H., Jacobson, S. W., et al. (2024). “Benchmark dose profiles for bivariate exposures”. In: *Risk Analysis* 44.10, pp. 2415–2428. DOI: 10.1111/risa.14303.
- Alexa, A. and Rahnenführer, J. (2024). “topGO: Enrichment Analysis for Gene Ontology”. In: R package version 2.56.0. DOI: 10.18129/B9.bioc.topGO.
- Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). “Improved scoring of functional groups from gene expression data by decorrelating GO graph structure”. In: *Bioinformatics* 22.13, pp. 1600–1607. DOI: 10.1093/bioinformatics/btl140.
- Altman, D. G. and Andersen, P. K. (1999). “Calculating the number needed to treat for trials where the outcome is time to an event”. In: *Bmj* 319, pp. 1492–1495. DOI: 10.1136/bmj.319.7223.1492.
- Andersen, J. S., Holst, H., Spliid, H., Andersen, H., Baun, A., and Nyholm, N. (1998). “Continuous ecotoxicological data evaluated relative to a control response”. In: *Journal of Agricultural, Biological, and Environmental Statistics*, pp. 405–420. DOI: 10.2307/1400573.
- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Corrected second printing, 1997. Springer New York, NY. DOI: 10.1007/978-1-4612-4348-9.
- Bornkamp, B., Pinheiro, J., and Bretz, F. (2009). “MCPMod: An R package for the design and analysis of dose-finding studies”. In: *Journal of Statistical Software* 29, pp. 1–23. DOI: 10.18637/jss.v029.i07.
- Bretz, F., Dette, H., and Pinheiro, J.C. (2010). “Practical considerations for optimal designs in clinical dose finding studies”. In: *Statistics in medicine* 29.7-8, pp. 731–742. DOI: 10.1002/sim.3802.
- Bretz, F., Hsu, J., Pinheiro, J., and Liu, Y. (2008). “Dose finding—a challenge in statistics”. In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50.4, pp. 480–504. DOI: 10.1002/bimj.200810438.
- Bretz, F., Pinheiro, J. C, and Branson, M. (2005). “Combining multiple comparisons and modeling techniques in dose-response studies”. In: *Biometrics* 61.3, pp. 738–748. DOI: 10.1111/j.1541-0420.2005.00344.x.
- Buckley, B. E., Piegorsch, W. W., and West, R. W. (2009). “Confidence limits on one-stage model parameters in benchmark risk assessment”. In: *Environmental and Ecological Statistics* 16.1, pp. 53–62. DOI: 10.1007/s10651-007-0076-2.
- Calderazzo, S., Tavel, D., Zurich, M.-G., and Kopp-Schneider, A. (2019). “Model-based estimation of lowest observed effect concentration from replicate experiments to identify potential biomarkers of in vitro neurotoxicity”. In: *Archives of toxicology* 93.9, pp. 2635–2644. DOI: 10.1007/s00204-019-02520-8.
- Carpenter, J. and Bithell, J. (2000). “Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians”. In: *Statistics in medicine* 19.9, pp. 1141–

1164. DOI: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.
- Chuang-Stein, C. and Agresti, A. (1997). "Tutorial in biostatistics a review of tests for detecting a monotone dose-response relationship with ordinal response data". In: *Statistics in medicine* 16.22, pp. 2599–2618. DOI: 10.1002/(SICI)1097-0258(19971130)16:22<2599::AID-SIM734>3.0.CO;2-9.
- Cole, T. J. and Green, P. J. (1992). "Smoothing reference centile curves: the LMS method and penalized likelihood". In: *Statistics in medicine* 11.10, pp. 1305–1319. DOI: 10.1002/sim.4780111005.
- Cox, D. R. (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.
- Cox, D. R. and Reid, N. (1987). "Parameter orthogonality and approximate conditional inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 49.1, pp. 1–18. DOI: 10.1111/j.2517-6161.1987.tb01422.x.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. 1. Cambridge university press. DOI: 10.1017/CB09780511802843.
- Delignette-Muller, M.-L., Forfait, C., Billoir, E., and Charles, S. (2011). "A new perspective on the Dunnett procedure: Filling the gap between NOEC/LOEC and ECx concepts". In: *Environmental Toxicology and Chemistry* 30.12, pp. 2888–2891. DOI: 10.1002/etc.686.
- Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.
- Deutsch, R. C. and Piegorsch, W. W. (2012). "Benchmark dose profiles for joint-action quantal data in quantitative risk assessment". In: *Biometrics* 68.4, pp. 1313–1322. DOI: 10.1111/j.1541-0420.2012.01811.x.
- Deutsch, R. C. and Piegorsch, W. W. (2013). "Benchmark dose profiles for joint-action continuous data in quantitative risk assessment". In: *Biometrical Journal* 55.5, pp. 741–754. DOI: 10.1002/bimj.201300037.
- Duda, J. C., Hengstler, J.G., and Rahnenführer, J. (2022a). "td2pLL: An intuitive time-dose-response model for cytotoxicity data with varying exposure durations". In: *Computational Toxicology* 23, p. 100234. DOI: <https://doi.org/10.1016/j.comtox.2022.100234>.
- Duda, J. C., Kappenberg, F., and Rahnenführer, J. (2022b). "Model selection characteristics when using MCP-Mod for dose-response gene expression data". In: *Biometrical Journal* 64.5, pp. 883–897. DOI: 10.1002/bimj.202000250.
- Efron, B. (1981). "Censored Data and the Bootstrap". In: *Journal of the American Statistical Association* 76.374, pp. 312–319. DOI: 10.1080/01621459.1981.10477650.
- Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. 1st ed. CRC Press. DOI: 10.1201/9780429246593.
- Eilers, P. H. C. and Marx, B. D. (1996). "Flexible smoothing with B-splines and penalties". In: *Statistical science* 11.2, pp. 89–121. DOI: 10.1214/ss/1038425655.

- Eilers, P. H. C. and Marx, B. D. (2002). “Generalized linear additive smooth structures”. In: *Journal of computational and graphical statistics* 11.4, pp. 758–783. DOI: 10.1198/106186002844.
- Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2022). *Regression: Models, Methods and Applications*. 2nd. Springer Berlin, Heidelberg. DOI: 10.1007/978-3-662-63882-8.
- Fahrmeir, L. and Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*. 2nd. Springer. DOI: 10.1007/978-1-4757-3454-6.
- Focke, W. W., Van der Westhuizen, I., Musee, N., and Loots, M. T. (2017). “Kinetic interpretation of log-logistic dose-time response curves”. In: *Scientific Reports* 7.1, p. 2234. DOI: 10.1038/s41598-017-02474-w.
- Ghallab, A., Myllys, M., Friebel, A., Duda, J. C., Edlund, K., Halilbasic, E., Vucur, M., Hobloss, Z., Brackhagen, L., Begher-Tibbe, B., Hassan, R., Burke, M., Genc, E., Frohwein, L. J., Hofmann, U., Holland, C. H., González, D., Keller, M., Seddek, A., Abbas, T., Mohammed, E. S. I., Teufel, A., Itzel, T., Metzler, S., Marchan, R., Cadenas, C., Watzl, C., Nitsche, M. A., Kappenberg, F., Luedde, T., Longerich, T., Rahnenführer, J., Hoehme, S., Trauner, M., and Hengstler, J. G. (2021). “Spatio-Temporal Multiscale Analysis of Western Diet-Fed Mice Reveals a Translationally Relevant Sequence of Events during NAFLD Progression”. In: *Cells* 10.10. DOI: 10.3390/cells10102516.
- González-Márquez, R., Schmidt, L., Schmidt, B. M., Berens, P., and Kobak, D. (2024). “The landscape of biomedical research”. In: *Patterns* 5.6. DOI: 10.1016/j.patter.2024.100968.
- Greene, W. H. (2018). *Econometric analysis*. 8th. Prentice Hall.
- Greenwood, M. (1926). “A Report on the Natural Duration of Cancer”. In: *Reports on Public Health and Medical Subjects* 33.
- Gu, X., Albrecht, W., Edlund, K., Kappenberg, F., Rahnenführer, J., Leist, M., Moritz, W., Godoy, P., Cadenas, C., Marchan, R., et al. (2018). “Relevance of the incubation period in cytotoxicity testing with primary human hepatocytes”. In: *Archives of toxicology* 92.12, pp. 3505–3515. DOI: 10.1007/s00204-018-2302-0.
- Hagemann, N. and Möllenhoff, K. (2025). “Overcoming Model Uncertainty—How Equivalence Tests Can Benefit From Model Averaging”. In: *Statistics in Medicine* 44.6, e10309. DOI: 10.1002/sim.10309.
- Hastie, T. and Tibshirani, R. (1986). “Generalized additive models”. In: *Statistical science* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604.
- Hernán, M. A. (2010). “The hazards of hazard ratios”. In: *Epidemiology* 21.1. Erratum in: *Epidemiology*, 22(1):134, 2011, pp. 13–15. DOI: 10.1097/EDE.0b013e3181c1ea43.
- Hothorn, L. (2014). “Statistical evaluation of toxicological bioassays - a review”. In: *Toxicology Research* 3.6, pp. 418–432. DOI: 10.1039/c4tx00047a.
- Hothorn, L. (2016). *Statistics in toxicology using R*. CRC Press Boca Raton, FL.
- Jachno, K., Heritier, S., and Wolfe, R. (2019). “Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice”. In: *BMC Medical Research Methodology* 19.103. DOI: 10.1186/s12874-019-0749-1.

- Jennrich, R. I. (1969). “Asymptotic properties of non-linear least squares estimators”. In: *The Annals of Mathematical Statistics* 40.2, pp. 633–643. DOI: 10.1214/aoms/1177697731.
- Jensen, S. M., Kluxen, F. M., and Ritz, C. (2019). “A review of recent advances in benchmark dose methodology”. In: *Risk Analysis* 39.10, pp. 2295–2315. DOI: 10.1111/risa.13324.
- Jensen, S. M. and Ritz, C. (2015). “Simultaneous inference for model averaging of derived parameters”. In: *Risk Analysis* 35.1, pp. 68–76. DOI: 10.1111/risa.12242.
- Jiang, X. and Kopp-Schneider, A. (2014). “Summarizing EC50 estimates from multiple dose-response experiments: A comparison of a meta-analysis strategy to a mixed-effects model approach”. In: *Biometrical Journal* 56.3, pp. 493–512. DOI: 10.1002/bimj.201300123.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. 2nd. Wiley Series in Probability and Statistics. John Wiley & Sons. DOI: 10.1002/9781118032985.
- Kang, S.-H., Kodell, R. L., and Chen, J. J. (2000). “Incorporating model uncertainties along with data uncertainties in microbial risk assessment”. In: *Regulatory Toxicology and Pharmacology* 32.1, pp. 68–72. DOI: 10.1006/rtph.2000.1404.
- Kaplan, E. L. and Meier, P. (1958). “Nonparametric estimation from incomplete observations”. In: *Journal of the American Statistical Association* 53.282, pp. 457–481. DOI: 10.2307/2281868.
- Kappenberg, F., Duda, J. C., Schürmeyer, L., Gül, O., Brecklinghaus, T., Hengstler, J. G., Schorning, K., and Rahnenführer, J. (2023). “Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review”. In: *Archives of Toxicology* 97.10, pp. 2741–2761. DOI: 10.1007/s00204-023-03561-w.
- Kappenberg, F., Grinberg, M., Jiang, X., Kopp-Schneider, A., Hengstler, J.G., and Rahnenführer, J. (2021). “Comparison of observation-based and model-based identification of alert concentrations from concentration-expression data”. In: *Bioinformatics* 37.14, pp. 1990–1996. DOI: 10.1093/bioinformatics/btab043.
- Kelly, C. and Rice, J. (1990). “Monotone smoothing with application to dose-response curves and the assessment of synergism”. In: *Biometrics*, pp. 1071–1085. DOI: 10.2307/2532449.
- Kerns, L. X. (2017). “Simultaneous confidence bands for log-logistic regression with applications in risk assessment”. In: *Biometrical Journal* 59.3, pp. 420–429. DOI: 10.1002/bimj.201600164.
- Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd. Statistics for Biology and Health. Springer New York, NY. DOI: 10.1007/b97377.
- Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis: A Self-Learning Text*. Ed. by M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, and W. Wong. 3rd. Statistics for Biology and Health. Springer New York, NY. DOI: 10.1007/978-1-4419-6646-9.

- Knol, M. J., Duijnhoven, R. G., Grobbee, D. E., Moons, K. G. M., and Groenwold, R. H. H. (2011). “Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials”. In: *PLoS One* 6.6, e21248. DOI: 10.1371/journal.pone.0021248.
- Kock, L. and Klein, N. (2025). “Truly multivariate structured additive distributional regression”. In: *Journal of Computational and Graphical Statistics* 34.4, pp. 1189–1201. DOI: 10.1080/10618600.2024.2434181.
- Kuss, O. and Hoyer, A. (2021). “A proportional risk model for time-to-event analysis in randomized controlled trials”. In: *Statistical Methods in Medical Research* 30.2, pp. 411–424. DOI: 10.1177/0962280220953599.
- Lange, M.R. and Schmidli, H. (2015). “Analysis of clinical trials with biologics using dose–time-response models”. In: *Statistics in Medicine* 34.22, pp. 3017–3028. DOI: 10.1002/sim.6551.
- Lee, K. J., Moreno-Betancur, M., Kasza, J., Marschner, I. C., Barnett, A. G., and Carlin, J. B. (2019). “Biostatistics: a fundamental discipline at the core of modern health data science”. In: *Medical Journal of Australia* 211.10, 444–446.e1. DOI: 10.5694/mja.50372.
- Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijlens, L. (2012). *Modeling dose-response microarray data in early drug development experiments using R: order-restricted analysis of microarray data*. Springer. DOI: 10.1007/978-3-642-24007-2.
- Loh, P.-L. (2021). “Searching for structure in complex data: a modern statistical quest”. In: DOI: 10.14760/SNAP-2021-003-EN.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd. Routledge. DOI: 10.1201/9780203753736.
- Mishra, P., Pandey, C. M., Singh, U., Keshri, A., and Sabaretnam, M. (2019). “Selection of appropriate statistical methods for data analysis”. In: *Annals of cardiac anaesthesia* 22.3, pp. 297–301. DOI: 10.4103/aca.ACA_248_18.
- Mokhtari, R. B., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017). “Combination therapy in combating cancer”. In: *Oncotarget* 8.23, p. 38022. DOI: 10.18632/oncotarget.16723.
- Möllenhoff, K., Dette, H., Kotzagiorgis, E., Volgushev, S., and Collignon, O. (2018). “Regulatory assessment of drug dissolution profiles comparability via maximum deviation”. In: *Statistics in Medicine* 37.20, pp. 2968–2981. DOI: 10.1002/sim.7689.
- Möllenhoff, K., Schorning, K., and Kappenberg, F. (Nov. 2022). “Identifying Alert Concentrations Using a Model-Based Bootstrap Approach”. In: *Biometrics* 79.3, pp. 2076–2088. DOI: 10.1111/biom.13799.
- Murado, M. A. and Prieto, M. A. (2013). “NOEC and LOEC as merely concessive expedients: Two unambiguous alternatives and some criteria to maximize the efficiency of dose–response experimental designs”. In: *Science of the total environment* 461, pp. 576–586. DOI: 10.1016/j.scitotenv.2013.04.098.

- Nelder, J. A. and Wedderburn, R. W. M. (1972). “Generalized linear models”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3, pp. 370–384. DOI: 10.2307/2344614.
- Oehlert, G. W. (1992). “A Note on the Delta Method”. In: *The American Statistician* 46.1, pp. 27–29. DOI: 10.2307/2684406.
- Pinheiro, J., Bornkamp, B., Glimm, E., and Bretz, F. (2014). “Model-based dose finding under model uncertainty using general parametric models”. In: *Statistics in medicine* 33.10, pp. 1646–1661. DOI: 10.1002/sim.6052.
- Rigby, R. A. and Stasinopoulos, M. D. (1996). “A semi-parametric additive model for variance heterogeneity”. In: *Statistics and Computing* 6.1, pp. 57–65. DOI: 10.1007/BF00161574.
- Rigby, R. A. and Stasinopoulos, M. D. (2005). “Generalized additive models for location, scale and shape”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 54.3, pp. 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.
- Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Chapman and Hall/CRC. DOI: 10.1201/9780429298547.
- Ritz, C., Baty, F., Streibig, J. C., and Gerhard, D. (2015). “Dose-response analysis using R”. In: *PloS one* 10.12, e0146021. DOI: 10.1371/journal.pone.0146021.
- Ritz, C., Jensen, S. M., Gerhard, D., and Streibig, J. C. (2019). *Dose-Response Analysis Using R*. Chapman and Hall/CRC. DOI: 10.1201/b21966.
- Ritz, C. and Streibig, J. C. (2008). *Nonlinear regression with R*. Springer. DOI: 10.1007/978-0-387-09616-2.
- Ritz, C. and Streibig, J. C. (2014). “From additivity to synergism—A modelling perspective”. In: *Synergy* 1.1, pp. 22–29. DOI: 10.1016/j.synres.2014.07.010.
- Rozman, K. K. and Doull, J. (2000). “Dose and time as variables of toxicity”. In: *Toxicology* 144.1-3, pp. 169–178. DOI: 10.1016/S0300-483X(99)00204-8.
- Ruberg, S. J. (1995a). “Dose response studies I. Some design considerations”. In: *Journal of Biopharmaceutical Statistics* 5.1, pp. 1–14. DOI: 10.1080/10543409508835096.
- Ruberg, S. J. (1995b). “Dose response studies II. Analysis and interpretation”. In: *Journal of biopharmaceutical statistics* 5.1, pp. 15–42. DOI: 10.1080/10543409508835097.
- Sand, S., D., Dietrich, Victorin, K., and A., Agneta (2006). “Identification of a critical dose level for risk assessment: developments in benchmark dose analysis of continuous endpoints”. In: *Toxicological Sciences* 90.1, pp. 241–251. DOI: 10.1093/toxsci/kfj057.
- Sauerbrei, W. and Buss, P. (2026). *STRATOS Initiative*. Accessed April 18, 2026. URL: <https://www.stratos-initiative.org/>.
- Schürmeyer, L., Sandig, L., Hezler, L. T., Igl, B.-W., and Schorning, K. (2025). “Optimal designs for identifying effective doses in drug combination studies”. In: *arXiv preprint*. DOI: 10.48550/arXiv.2506.05913.
- Schüttler, A., Altenburger, R., Ammar, M., Bader-Blukott, M., Jakobs, G., Knapp, J., Krüger, J., Reiche, K., Wu, G., and Busch, W. (2019). “Map and model—moving

- from observation to prediction in toxicogenomics”. In: *GigaScience* 8.6, giz057. DOI: 10.1093/gigascience/giz057.
- Schwartz, L. M., Woloshin, S., and Welch, H. G. (1999). “Misunderstandings about the effects of race and sex on physicians’ referrals for cardiac catheterization”. In: *New England Journal of Medicine* 341.4, pp. 279–283. DOI: 10.1056/NEJM199907223410411.
- Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear regression*. New Jersey: John Wiley & Sons. DOI: 10.1002/0471725315.
- Serra, A., Fratello, M., Del Giudice, G., Saarimäki, L. A., Paci, M., Federico, A., and Greco, D. (2020). “TinderMIX: Time-dose integrated modelling of toxicogenomics data”. In: *Gigascience* 9.5, giaa055. DOI: 10.1093/gigascience/giaa055.
- Slikker Jr., W., Andersen, M. E., Bogdanffy, M. S., Bus, J. S., Cohen, S. D., Conolly, R. B., David, R. M., Doerrer, N. G., Dorman, D. C., Gaylor, D. W., et al. (2004). “Dose-dependent transitions in mechanisms of toxicity”. In: *Toxicology and Applied Pharmacology* 201.3, pp. 203–225. DOI: 10.1016/j.taap.2004.06.019.
- Staniswalis, J. G. and Cooper, V. (1988). “Kernel estimates of dose response”. In: *Biometrics*, pp. 1103–1119. DOI: 10.2307/2531739.
- Stasinopoulos, M. D., Kneib, T., Klein, N., Mayr, A., and Heller, G. Z. (2024). *Generalized additive models for location, scale and shape: a distributional regression approach, with applications*. Vol. 56. Cambridge University Press. DOI: 10.1017/9781009410076.
- Stasinopoulos, M. D. and Rigby, R. A. (2008). “Generalized additive models for location scale and shape (GAMLSS) in R”. In: *Journal of Statistical Software* 23, pp. 1–46. DOI: 10.18637/jss.v023.i07.
- Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman and Hall/CRC. DOI: 10.1201/b21973.
- Sutradhar, R. and Austin, P. C. (2018). “Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios”. In: *Annals of Epidemiology* 28.1, pp. 54–57. DOI: 10.1016/j.annepidem.2017.10.014.
- The Academy of Medical Sciences (2017). *Sources of Evidence for Assessing the Safety, Efficacy and Effectiveness of Medicines*. <https://acmedsci.ac.uk/file-download/86466482> (accessed 06 March 2023).
- Van Der Hoeven, N. (1997). “How to measure no effect. Part III: Statistical aspects of NOEC, ECx and NEC estimates”. In: *Environmetrics: The Official Journal of the International Environmetrics Society* 8.3, pp. 255–261. DOI: 10.1002/(SICI)1099-095X(199705)8:3<255::AID-ENV246>3.0.CO;2-P.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/CB09780511802256.
- Warne, Michael St J and Van Dam, Rick (2008). “NOEC and LOEC data should no longer be generated or used”. In: *Australasian Journal of Ecotoxicology* 14.1, pp. 1–5.
- West, R. W. and Kodell, R. L. (1993). “Statistical methods of risk assessment for continuous variables”. In: *Communications in Statistics-Theory and Methods* 22.12, pp. 3363–3376. DOI: 10.1080/03610929308831221.

- Wood, S. N. (2017). *Generalized additive models: an introduction with R*. 2nd. chapman and hall/CRC. DOI: 10.1201/9781315370279.
- Wu, Y., Piegorsch, W. W., West, R. W., Tang, D., Petkewich, M. O., and Pan, W. (2006). “Multiplicity-adjusted inferences in risk assessment: benchmark analysis with continuous response data”. In: *Environmental and Ecological Statistics* 13.1, pp. 125–141. DOI: 10.1007/s10651-005-5695-x.
- Yuan, Y. and Yin, G. (2011). “Dose–response curve estimation: a semiparametric mixture approach”. In: *Biometrics* 67.4, pp. 1543–1554. DOI: 10.1111/j.1541-0420.2011.01620.x.
- Zhang, Z., Ambroggi, F., Bokov, A. F., Gu, H., Beurs, E. de, and Eskaf, K. (2018). “Estimate risk difference and number needed to treat in survival analysis”. In: *Annals of translational medicine* 6.7, p. 120. DOI: 10.21037/atm.2018.01.36.
- Zhou, Y., Sloan, A., Menon, S., and Wang, L. (2024). “Combination MCP-Mod for two-drug combination dose-ranging studies”. In: *Journal of Biopharmaceutical Statistics* 35.2, pp. 257–270. DOI: 10.1080/10543406.2024.2311254.
- Zhu, Y. (2005). “Dose–time–response modeling of longitudinal measurements for neurotoxicity risk assessment”. In: *Environmetrics: The official journal of the International Environmetrics Society* 16.6, pp. 603–617. DOI: 10.1002/env.725.

Further publications

Ghallab, A., ..., Ameis, L. ..., & Hengstler, J. G. (2025). Stage-dependent effects of systemic ASBT inhibition in a cholestasis-induced cholemic nephropathy mouse model. *JHEP Reports*, 101599. DOI: 10.1016/j.jhepr.2025.101599

Contribution: Bioinformatical analysis